Novel Computational Tools for Linking Genotypes to Microbial Community Phenotypes

---

A Dissertation

Presented to

the faculty of the School of Engineering and Applied Science

University of Virginia

---

in partial fulfillment

of the requirements for the degree

Doctor of Philosophy

by

Matthew Bryon Biggs

December

2016

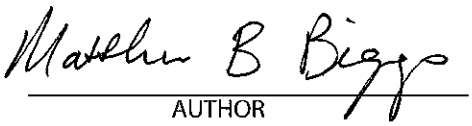
APPROVAL SHEET

The dissertation

is submitted in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy



AUTHOR

The dissertation has been read and approved by the examining committee:

Jason A. Papin, PhD

Advisor

Jeffrey J. Saucerman, PhD

Peter M. Kasson, MD, PhD

Martin Wu, PhD

Richard L. Guerrant, MD

Accepted for the School of Engineering and Applied Science:

Craig H. Benson, Dean, School of Engineering and Applied Science

December

2016

Contents

List of Figures

List of Tables

Acknowledgments

I find it difficult to imagine a better graduate school experience than mine has been, and all of the credit goes to the many wonderful people in my life.

First of all, I thank my wife Jessica. She has offered unfailing encouragement through all the highs and lows of classes and research. She has offered warm companionship, loving friendship, clear-headed advice, and always helps me to have healthy priorities and balance. I dedicate this dissertation to her, because in many respects, this is as much a product of her efforts and perseverance as it is mine. I want to thank our children—John and Lucy—who serve as my biggest motivation to succeed, and who are an ever-present source of adventure.

I thank my parents who provided me with an early love of learning. They taught me to read (which has been really handy as a graduate student), to tinker, and to communicate. Possibly even more valuable, they taught me by example (and by giving me the opportunity) to be self-motivated and to follow my interests. Graduate school was in many respects, just an extension of my upbringing in the Biggs home. I am grateful for their continuing support, examples of lifelong learning, and good priorities. Similarly, I thank my siblings for enriching my life through their humor and wide-ranging interests and experiences.

I count myself extremely lucky to have had Jason Papin as my advisor, mentor and friend. His unselfish desire to build students was apparent to me early on, when, as an undergraduate at BYU, I contacted him out of the blue to ask for advice on how to start an iGEM team. Not knowing me, he still gave time for several phone conversations to give me advice about everything from iGEM to science careers. Throughout my time as a graduate student in his lab, he has encouraged initiative and big ideas, has gently coaxed me through discouragement, self-doubt and shortsightedness. He is an outstanding example of optimism, continuous curiosity, life balance, and selflessness, and I aspire to follow his example.

I am grateful for the influence of many others. My fellow students in the lab have always provided friendship and a supportive social environment that makes "working" in the lab feel like anything but work. I thank the students who paved the way for me and taught me through word and example: Paul Jensen, Jennie Bartell, Kevin D'Auria, and Edik Blais. I thank them for their ongoing friendship and mentoring. Many thanks to the current members of the Papin lab who have enriched my experience with their generosity and friendship: Phil Yen, Anna Blazier, Greg Medlock, Kris Rawls, Maureen Carey, Tom Moutinho, Bonnie Dougherty and Laura Dunphy. I thank them for all the welcomed distractions and shenanigans, sincere feedback, examples of hard work and creativity, and the constant supply of baked goods. I am indebted to Glynis Kolling, who spent countless hours discussing the details of lab protocols, asking insightful questions about my projects, giving career advice, and sharing ideas for weekend cooking projects. Thanks to my grandfather, Cam Mosher, the first PhD in my family tree. I am grateful that he has nurtured my scientific curiosity, critical thinking and entrepreneurial spirit for as long as I can remember. I express my love and appreciation to Jessica's parents, Michael and Pamela Kosorok, for their trust in me, their good advice and willingness to join in all my crazy project ideas. I am thankful for my mentors at BYU who gave me my first training as a scientist and set my course for graduate school, specifically Joel Griffitts and Julianne Grose. I also want to thank Maia Donahue for her mentoring during my internship at Dow Agrosciences, and for her continued support as I move forward in my career.

Many other people have positively influenced my graduate school experience, too many to mention them all by name. Thanks go to my professors, undergraduates who I had the privilege of working with, and to collaborators here and at other institutions. Thank you to my dissertation committee for the many fruitful meetings, both as a group and one-on-one in which you have given helpful feedback, advice and encouragement.

To all of you, my sincerest thanks!

Chapter 1

Introduction

1.1 Specific Aims

The human microbiome is vital to human health as a metabolic "organ" with important catabolic and anabolic functions. Development of microbiome-targeted therapies requires mechanistic understanding of how microbes interact with each other and with the host. Mechanistic computational approaches can increase knowledge gains from experiments, infer system parameters that cannot be measured, and accelerate the design of novel therapies. Constraint-based reconstruction and analysis (COBRA) of genome-scale metabolic networks is a powerful mechanistic approach that has been widely applied to the study of metabolism in single species, but is underdeveloped and underutilized as a tool for studying metabolism in microbial communities. The application of COBRA methods to microbial communities is still hampered by:

1. A lack of approaches for integrating metabolic networks with other community data such as spatial organization or known interactions

2. The difficulty of characterizing individual species in a complex community

3. The infeasibility of manually reconstructing metabolic network for hundreds of species in a community

To address these challenges, we completed the following objectives:

Aim 1: Integrate metabolic networks into multiscale models of microbial communities. Within microbial communities, metabolism influences and is influenced by other physical and chemical factors such as spatial structure and interspecies interactions. Until our work, genome-scale metabolic networks were used primarily to model well-mixed communities. For the first part of this objective, we created the first (to our knowledge) multiscale model that integrates metabolism with spatial distribution and we recapitulated known features of *Pseudomonas aeruginosa* biofilm formation. For the second part of this objective, we integrated a Boolean network representation of interspecies interactions in a murine microbiome with metabolic networks to explore the role of metabolism in species interaction networks.

Aim 2: Reconstruct species-level metabolic networks from metagenomic samples. Attempts to model the metabolism of microbial communities are often limited by a lack of knowledge about the constituent species of the communities. Culture-free techniques such as metagenomic sequencing provide information about the genomes of all species in a community, but the genomic fragments are mixed together. A challenge in the field is to sort mixed sequence fragments into the correct species bins, and we found that metabolic networks can guide the binning process. We developed the SOrting by NEtwork Completion (SONEC) approach, which improves the sequence coverage of individual species' genomes and simultaneously reconstructs a metabolic network for each species. The SONEC approach improves access to species-specific information which can then be used in modeling approaches.

Aim 3: Improve predictions generated by draft metabolic networks. Metabolic network reconstructions require months of manual curation before they can be used to make useful predictions. Because the human microbiome contains hundreds of different species which vary from person to person and across time, it is infeasible to manually curate models for each individual species. To address this problem, we developed an ensemble approach, which improves the prediction accuracy of automatically-generated networks by pooling predictions from several draft networks representing the same species. Each network within the ensemble represents a hypothesized network structure, and these are reconstructed based on random subsets and permutations of the available data for each species. We found that ensemble predictions are superior to those of single reconstructions with respect both to overall network properties (such as nutrient utilization predictions) and to mechanistic details (such as gene essentiality). Our ensemble approach leverages the speed of automated network reconstruction while improving predictions, making it possible to generate valuable predictions about complex microbial communities within a practical time span.

Multiscale models which include metabolic network reconstructions have the potential to generate detailed hypotheses linking genes to community phenotypes. In completing these aims, we developed two novel multiscale models which integrated metabolic network mod-

els with spatial structure and interaction networks. Furthermore, we overcame two substantial obstacles hindering the future application of such multiscale models by improving the process of reconstructing genomes from metagenomic sequence fragments, and by improving the accuracy of draft models using an ensemble approach. The completion of these objectives paves the way for much broader adoption of mechanistic, multiscale models in microbiome research.

1.2 A Preview of this Dissertation

I chose to organize the specific aims in the chronological order in which the work occurred. I hope that in presenting my work this way it is clear to the reader how the inspiration for the second and third aims arose naturally from the first. Aim 1 consists of the earliest work where I ambitiously jumped ahead to address the engineering problem of multiscale modeling of microbial communities. In the process of developing such models, I discovered substantial barriers to implementing those multiscale models in the real-life setting of complex microbial communities. Aim 2 addresses one barrier, which is the lack of reference genomes for most members of natural microbial communities. Aim 2 addresses a second barrier, which is the need for substantial manual curation of metabolic networks.

The organization of this dissertation follows the three specific aims above, with some additional material at the end to provide context for future directions. My work relating to Aim 1 is found in Chapters 2–4. Chapter 2 describes at a multiscale model of *P. aeruginosa* biofilm formation. Chapter 3 describes a multi-faceted analysis of a gut microbial community using both dynamic network analysis and genome-scale metabolic networks. Chapter 4 is a review article describing the available techniques for modeling microbial interactions using genome-scale metabolic networks and future directions for the field. My work relating to Aim 2 is found in Chapter 5, where I describe and demonstrate the SONEC algorithm. My work relating to Aim 3 is found in Chapter 6, where I present a novel approach to modeling microbial metabolism using ensembles of genome-scale metabolic networks.

In addition to completing work on the three aims above, I completed additional experimental work that will soon serve a powerful role in improving our ability to test and validate the types of computational approaches I developed in Aims 1–3. In Chapter 7, I present the results of our genetic and metabolic characterization of a model murine gut microbiota known as the altered Schaedler flora (ASF). Finally, in Chapter 8 I discuss the impact of the work reported in this dissertation, and future directions.

Two helpful asides: One, despite having a focus on microbial communities, several of the computational projects make use of the single species *Pseudomonas aeruginosa*. This is because previous research in the Papin lab resulted in high quality metabolic network reconstructions of *P. aeruginosa* which is convenient for purposes of algorithm validation. Two, there are many mentions of "Supplemental Materials", all of which are available for each chapter through the online version of the corresponding published manuscript.

Chapter 2

Novel Multiscale Modeling Tool Applied to *Pseudomonas aeruginosa* Biofilm Formation

The text for this chapter has been previously published as a research article here:

Biggs MB and Papin JA. (2013). Novel Multiscale Modeling Tool Applied to *Pseudomonas aeruginosa* Biofilm Formation. *PLOS One*, 8(10):e78011.

2.1 Context

This paper began as a class project in SYS6035: Agent-based Modeling and Simulation of Complex Systems. The instructor, Dr. Gerard Learmonth, encouraged us to make publication a goal of our experience in the class. I found that starting with publication as the end goal motivated my learning and my choice of class project. At the time of writing of this dissertation, this paper had been cited 13 times. I continue to receive personal inquiries through email about the MatNet tool described in this paper. I have corresponded with researchers who are using MatNet at many institutions in the United States and around the world, including the University of Michigan, the Massachusetts Institute of Technology, Ohio State University, the United States Department of Agriculture, and institutions in Iran, Israel, France, Germany, Austria, and the United Kingdom.

2.2 Synopsis

Multiscale modeling is used to represent biological systems with increasing frequency and success. Multiscale models are often hybrids of different modeling frameworks and programming languages. We present the MATLAB-NetLogo extension (MatNet) as a novel tool for multiscale modeling. We demonstrate the utility of the tool with a multiscale model of *Pseudomonas aeruginosa* biofilm formation that incorporates both an agent-based model (ABM) and constraint-based metabolic modeling. The hybrid model correctly recapitulates oxygen-limited biofilm metabolic activity and predicts increased growth rate via anaerobic respiration with the addition of nitrate to the growth media. In addition, a genome-wide survey of metabolic mutants and biofilm formation exemplifies the powerful analyses that are enabled by this computational modeling tool.

2.3 Introduction

Multiscale modeling is a broad class of hybrid modeling techniques that attempt to represent physical systems that span multiple spatial or time scales. Spatial and time scales are particularly interdependent in biological applications and there is increasing utility for multiscale models that capture this interdependency [1]. A recent example is a model of vascular adaptation that combines an agent-based model (i.e. cellular level) with a continuum biomechanical model (i.e. tissue level) [2, 3]. Using this model, Hayenga *et al.* identify causal factors in arterial adaptation to sustained increases in blood pressure. These predicted factors are active at different spatial scales and include cell growth and tissue remodeling. This remodeling in turn occurs as a function of the changes in production and removal of collagen and smooth muscle cells due to hemodynamically-induced stresses, emphasizing the highly multiscale nature of the biological system and the need for mathematical models that integrate data from disparate spatial and temporal scales [2]. Multiscale models show significant potential for representing the inherent complexity of biological systems, generating testable hypotheses to understand fundamental mechanisms.

The hybrid nature of many multiscale models creates a need for software tools in which to implement the models. Different software packages offer unique strengths (e.g. R provides vast statistics capabilities [4], NetLogo provides a rich environment for agent-based modeling [5], and MATLAB offers a wealth of engineering tools [6]). It is often advantageous to implement separate portions of a model in the most appropriate language and to combine the results dynamically. Dynamically combining model results between software platforms can be achieved with packages written for that purpose. Examples of current packages that perform this function

are the NetLogo-R extension by Thiele and Grimm [7] and R.matlab by Bengtsson [8]. As multiscale models are built with increasingly diverse computational components, more tools will be needed that facilitate dynamic integration of disparate software tools.

Here, we present a novel software tool that fills a need in biomedical and biological multiscale modeling. The MATLAB-NetLogo extension (MatNet) provides new functions within NetLogo that allow data passing between NetLogo and MATLAB, and the calling of any valid, one-line MATLAB commands from within NetLogo. The need for this tool is demonstrated by publications that have used NetLogo and MATLAB (as the most appropriate software platforms) to implement biomedical multiscale models [2, 3, 9]. The new tool presented herein facilitates future dynamic integration of these software platforms.

To demonstrate the utility of this tool, we present a multiscale model of *Pseudomonas aeruginosa* biofilm growth. *P. aeruginosa* is a common opportunistic pathogen that forms biofilms on medical implants [10] and in the lungs of cystic fibrosis patients [11], and is a model organism for biofilm formation. In our model, we combine an existing ABM of biofilm development [12, 13] with a genome-scale metabolic model of *P. aeruginosa* metabolism [14]. This biofilm model is multiscale in its incorporation of biofilm-level spatial information such as structural remodeling and nutrient diffusion, as well as cell-level details of metabolic functions such as nutrient uptake and growth yields. The ABM, originally developed in C++, was implemented in NetLogo to exploit the existing framework and flexibility it offers as an ABM platform [15]. Metabolic modeling was implemented in MATLAB as done previously [16]. The resulting model reproduces known biofilm structure from limited oxygen diffusion. The model further demonstrates the utility of MatNet by generating hypotheses for how gene-level perturbations influence biofilm structure.

2.4 Methods

Agent-Based Model of Biofilm Structure

Here, we briefly describe the structure and processes of the ABM and refer the reader to our publicly-available model as well as corresponding citations for further details. The rules for the two-dimensional ABM of biofilm growth were implemented in NetLogo essentially as described by Pizarro *et al.* [12, 13]. The purpose of the ABM is to capture emergent biofilm structure that results from growth and dispersion of individual bacterial cells. The biofilm is represented as a two-dimensional cross-section divided into squares. Each square represents a region of liquid growth media. As such, each square contains variables that represent nutrient levels in that area, and nutrients are allowed to diffuse from higher to lower concentrations. Each agent in the simulation represents bacteria. Agents diffuse randomly unless adjacent to "biofilm". "Biofilm" is defined in the simulation as agents directly adjacent to the bottom surface of the simulated space, or adjacent to a chain of agents that terminates at the bottom surface. Agents in the biofilm do not move except as a result of division. Bacterial agents undergo binary division once the nutrients consumed exceed a pre-defined threshold. Only one agent may occupy a square; therefore, once an agent divides into two, the new agent is placed in a randomly-selected adjacent square, and if that square is occupied, the next agent is displaced to a random adjacent square. This process, termed "shoving", is continued until no square contains more than one agent.

The key difference in our model from the Pizarro *et al.* formulation is a change from representative "food particles" to concentrations of all 105 extracellular metabolites used in the genome-scale metabolic network reconstruction of *P. aeruginosa* [14]. Each metabolite diffuses independently as a function of the molecular mass. Metabolites diffuse more slowly through regions of the ABM space defined as biofilm (60% of aqueous rate for gases, and 25% of aqueous rate for all other metabolites) [17].

The multiscale modeling of the biofilm is an iterative process involving analysis in MATLAB and NetLogo. First, constraints on exchange fluxes for the FBA problem in MATLAB are scaled to local nutrient concentrations. This simplifying assumption can be relaxed with more detailed flux constraints implemented as such data is available. However, these simplified constraints are sufficient to illustrate the value of the modeling tool presented here. After solving the FBA problem in MATLAB, local nutrient concentrations are calculated and returned, along with the growth rate, to the NetLogo environment. The nutrient concentrations are updated in NetLogo, agents with accumulated biomass divide in two and rearrange according to the shoving rule, nutrients diffuse, and the new nutrient concentrations are passed to MATLAB. These steps constitute one time step of the simulation, which simulates a 5 minute interval of biofilm growth. A single simulation of 200 time steps simulates biofilm growth over a period of $\sim$17 hours.

Our implementation of the biofilm model in NetLogo displays the same behavior as the Pizarro *et al.* model (Figure S1). Because the ABM was independently validated previously [12, 13], it will not be further validated here except as pertains to the hybrid metabolic and agent-based models.

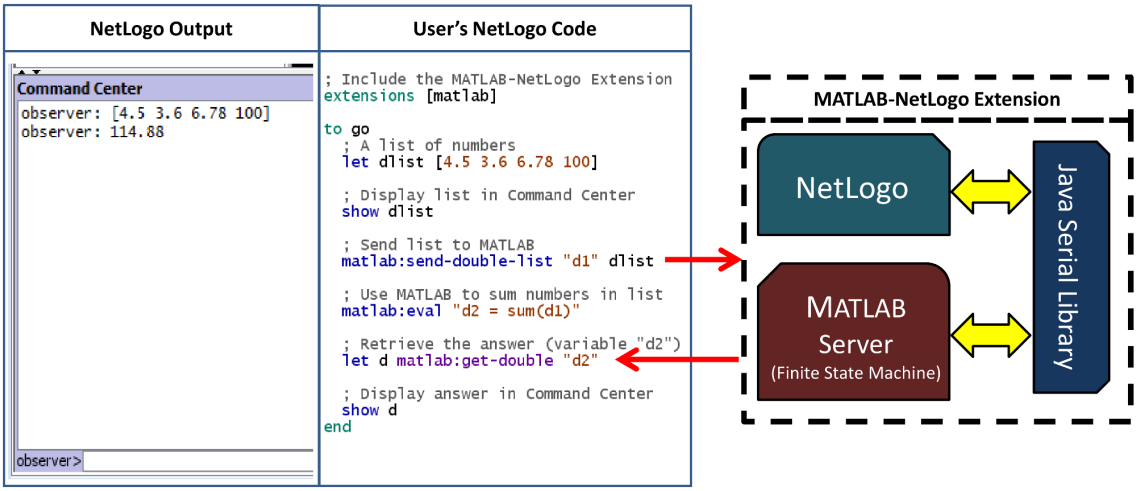


Figure 2.1: **MATLAB-NetLogo Extension (MatNet) diagram and example code.** MATLAB and NetLogo are both Java-based applications and are able to pass data via the Java Serial library. The user is insulated from the details of data passing, and can call MATLAB functions (native or user-defined) from within NetLogo using simple commands. In the example above, a list of numbers is created in NetLogo and passed to MATLAB where the numbers are summed. The answer is retrieved from MATLAB and displayed in NetLogo.

Genome-Scale Metabolic Network Reconstruction

P. aeruginosa metabolism was modeled using the previously published genome-scale metabolic reconstruction [14]. The model was analyzed with functions from the COBRA Toolbox [18] implemented previously in MATLAB. The COBRA Toolbox utilized the Gurobi optimizer [19]. Metabolite concentrations in each occupied square of the ABM were used to constrain uptake rates in the model. Discrete solutions for each cell agent at each time point were found using flux balance analysis (FBA) [20]. Cell agent biomass and metabolite concentrations were updated using dynamic FBA [21].

Metabolic Model Constraints

Initial conditions simulating glucose minimal media were generated by including negative, non-zero lower bounds for the extracellular metabolite exchange reactions: Iron (Fe and Fe^{3+}), Oxygen (O_2), D-Glucose ($C_6H_{12}O_6$), Cadmium (Cd), Carbon Dioxide (CO_2), Sulfate (H_2O_4S), Copper (Cu), Water (H_2O), Manganese (Mn), Cobalt (Co), Ammonium (NH_4+), Sodium (Na), Nitrogen (N_2), Magnesium (Mg), Orthophosphate (H_3O_4P), and Zinc (Zn). For the anaerobic respiration simulation, an additional negative, non-zero lower bound was included for the Nitrate (HNO_3) exchange reaction. The metabolic model and accompanying constraints were previously described [14] and were not further validated here except as pertains to the hybrid model.

Software Availability

MatNet, example code, and the biofilm model are available from: bme.virginia.edu/csbl/Downloads1-matnet.html

Simulation Specifications

Simulations were performed on a 64-bit Sony Vaio laptop with 6 GB of RAM and a 2.8 GHz dual-core processor running Windows 7, NetLogo version 5.0.3 and MATLAB version 2012b. The duration of single simulations of biofilm growth ranged from 5 to 15 hours, depending on model settings.

2.5 Results/Discussion

Novel Multiscale Modeling Tool

MatNet was written in Java, utilizing the NetLogo Extensions API (Figure 2.1). NetLogo and MATLAB pass data using the Java Serial library. MATLAB is opened as a background process and runs a server script that is an implementation of a finite state machine. The architecture was based on R.matlab [8] and the NetLogo-R extension [7]. This extension adds nine commands or "primitives" for sharing and evaluating data with MATLAB from within NetLogo (see "User Guide" in Supplemental Material S1). The resulting extension provides a simple interface between the NetLogo and MATLAB platforms that allows users to exploit the strengths of both languages in their models (Figure 2.1). While the following multiscale analysis is a biomedical example, this tool could readily find application in other fields for

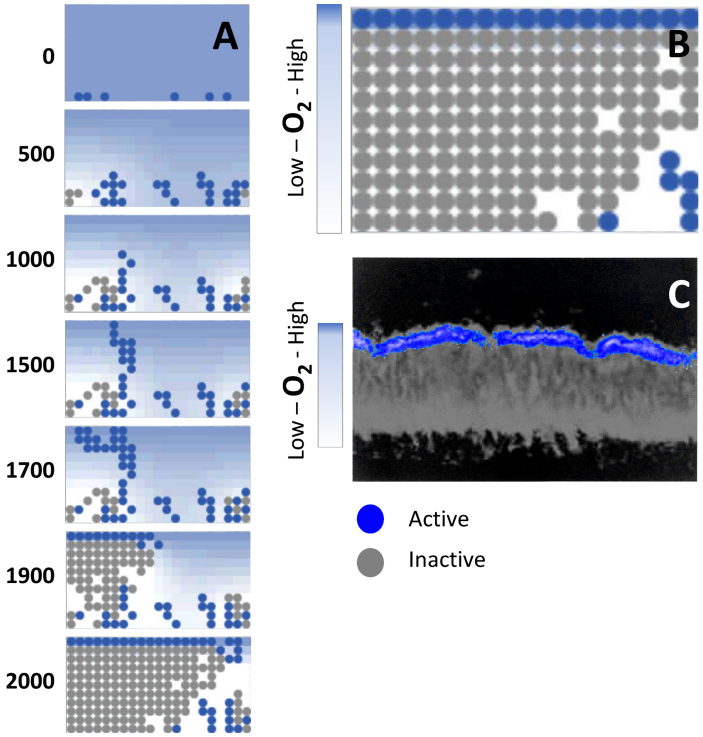


Figure 2.2: **Oxygen-dependent metabolic activity in *P. aeruginosa* biofilms.** (A) Progression of biofilm growth in a multiscale model with the associated time step (time steps represent 5 minute intervals). Each circle represents a cluster of *P. aeruginosa* cells. (B) Snapshot from multiscale biofilm model in glucose minimal media at time step 2,000. (C) *in vitro P. aeruginosa* biofilm cross section grown in glucose MOPS media for four days (modified from Xu *et al.* [22]). The oxygen gradient through the biofilm limits metabolic activity. Only with high O_2 (near the surface) can cells actively synthesize protein. The multiscale model recapitulates this pattern of oxygen-limited metabolic activity throughout the biofilm.

which integrated MATLAB and NetLogo analyses are of value such as ecology [23], finance [24], or behavioral science [25].

Individual simulations were performed over 5 to 15 hours. We evaluated the computational time for each of the functions in a given simulation. A large fraction of the simulation run time is claimed by the metabolite diffusion simulations in NetLogo and the repeated FBA simulations in MATLAB. The slower run time of these steps is expected, given that both processes are called frequently during each time step, and both are computationally intensive. While an appreciable portion of the computational time was spent passing data between MATLAB and NetLogo, this computational time is attributable to the high frequency with which these functions were called. The passing of data between the two environments via MatNet did not add undue computational overhead. Among all the functions in the simulation, each MatNet function was listed among the fastest on a per-function-call basis.

Oxygen-Limited Metabolic Activity in a *P. aeruginosa* Biofilm Model

The ABM correctly recapitulates oxygen-limited metabolic activity in a biofilm (Figure 2.2A). Biofilm formation was simulated under glucose minimal media conditions. Metabolic activity was defined as an increase in biomass (> 0.01 mass dry weight) associated with a particular agent in the two-dimensional space. Metabolites were allowed to diffuse in from the top to mimic fresh media being washed over the biofilm as done by Pizarro *et al.* [12, 13]. Oxygen at the top was held at a constant 0.21 mM [21]. All simulations showed reduced metabolic activity in the interior of the biofilm, and increased metabolic activity at the surface. An evaluation of the exchange reaction fluxes in the metabolic models indicated oxygen as the limiting metabolite (Figure 2.2B), consistent with findings from Xu *et al.* who report oxygen-limited growth in *P. aeruginosa* biofilms (Figure 2.2C) [22]. Furthermore, metabolic activity (as measured by protein synthesis) is restricted to a layer of cells at the biofilm surface (Figure 2.2C) as previously reported [22]. Therefore, this model of biofilm growth correctly recapitulated known characteristics of *P. aeruginosa* biofilm.

Nitrate Promotes Anaerobic Respiration and Increased Biofilm Growth Rate

Our multiscale model recapitulated increased biofilm growth rate in nitrate-supplemented media (Figure 2.3A). Addition of nitrate (NO_3) to the *in silico* growth media increased biofilm growth rate by approximately 10-fold, as determined by the change in cell agent counts over the first 263 time steps (Figure 2.3B). Nitrate relieves the oxygen limitation in *P. aeruginosa* by allowing anaerobic growth via denitrification [22, 26]. Denitrification, or anaerobic respiration, is the process whereby nitrate (NO_3) is reduced to dinitrogen (N_2), and nitrate replaces gaseous oxygen as the terminal electron acceptor. Anaerobic respiration prolongs active growth deeper in the biofilm after oxygen is removed from the microenvironment. The model prediction of increased growth rate was subsequently validated via literature search; Borriello *et al.* report increased biofilm growth with the addition of nitrate [27]. Although a direct comparison is not possible due to different growth conditions than those simulated in the model, the results reported by Borriello *et al.* serve as a qualitative validation for

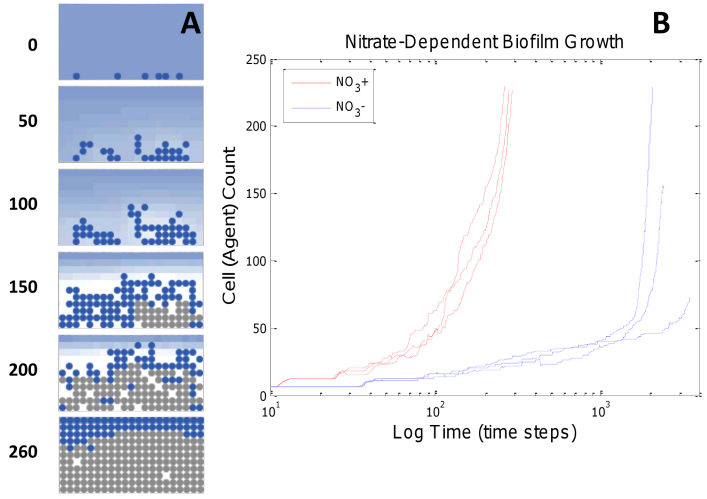


Figure 2.3: **ABM simulations of nitrate-dependent growth rates.** (A) Predicted biofilm formation in the presence of nitrate (NO_3) shows higher proportion of active cells when compared to glucose minimal media control (Figure 2.2). (B) Predicted biofilm growth with and without nitrate (3 independent runs each). Addition of nitrate is predicted to increase biofilm growth rate by enabling anaerobic growth deeper in the biofilm. Note that for simulations in glucose minimal media (blue lines), slower growth increases the impact of random cell spacing and resultant heterogeneous nutrient usage such that the model resulted in differing final cell counts for the same 15 hour simulation times.

the model predictions. This validated model prediction demonstrates that hybrid ABM-metabolic models can display predictive emergent behavior that is physiologically relevant.

in silico Gene-Deletion Screen

An *in silico* gene-deletion screen predicts the influence of individual genes on biofilm growth. Genes were deleted from the metabolic model by constraining reaction flux to zero. All possible single-gene deletions were evaluated in MATLAB using FBA. From the results of this analysis, a subset of metabolic models was selected to represent a range of growth phenotypes (lethal, sub-optimal and wild-type). A multiscale model was generated for each mutant background selected and was evaluated for 200 time steps on nitrate-supplemented glucose minimal media (Figure 2.4). Qualitative behavior was clearly evident by time step 200, which was chosen consequently as a stopping point. Note that with MatNet a genome-wide gene deletion screen and the resulting phenotypic differences of a multicellular system can quickly and easily be explored, thus providing useful hypotheses to guide experimental design.

We present the hybrid model results for nine models: wild-type, Δ*sdhD*, Δ*nasA*, Δ*gcd*, Δ*wbpL*, Δ*aceE*, Δ*pgm*, Δ*atpD*, and Δ*lysS*. The wild-type model served as a positive control, while Δ*lysS* served as a negative control (*lysS* encodes a tRNA synthetase and is an essential gene on nitrate-supplemented glucose minimal media). Reduced growth was predicted for Δ*sdhD*, Δ*aceE* and Δ*atpD*. *sdhD* plays a role in aerobic respiration [28] and its deletion restricts growth by limiting cells to anaerobic respiration. *atpD* encodes a subunit of ATP synthase. *aceE* encodes a pyruvate dehydrogenase and its deletion uncouples the citric acid cycle from glycolysis. Severely restricted growth (only slightly more biomass was found at time step 200 than what was initially seeded into the system) was predicted for Δ*gcd* and Δ*pgm*. *gcd* encodes a glucose dehydrogenase and de Werra *et al.* report that on glucose minimal media, mutant strains without *gcd* initially grow very slowly [29]. *pgm* encodes a phosphoglycerate mutase. The Δ*nasA* model is of interest because *nasA* encodes a nitrate transporter, and yet the model predicts near-wild-type growth on nitrate-supplemented media. Further investigation showed that the metabolic reconstruction contains two independent nitrate transport pathways. In the Δ*nasA* model, nitrate is taken into the cell via a separate nitrate ABC transporter encoded by PA2294, PA2295, PA2296, or PA2327, PA2328, PA2329. The results of the Δ*nasA* model are of further interest because they highlight the utility of this multiscale modeling approach to explore the interplay of gene function and biofilm microenvironment heterogeneity. While some model predictions were validated through literature search, the unsupported predictions stand as hypotheses awaiting experimental validation. The purpose of this screen is simply to demonstrate the power of our hybrid model to survey genome-wide, gene-level perturbations on biofilm-level phenotype.

2.6 Conclusion

This model framework correctly recapitulated known biofilm characteristics and yielded useful predictions that may guide future experimental design. Future development of the models presented here could include an accounting of extracellular polymeric substances in the ABM [30–33], the addition of rules linking specific genes to biofilm growth, and the inclusion of gene regulation in the metabolic model. Another potential biological process highly amenable to hybrid modeling using MatNet is quorum sensing, in which spatial information of the cells contributes to the signaling and gene regulation of the bacteria. Models of quorum sensing could also be integrated with the biofilm model, facilitating

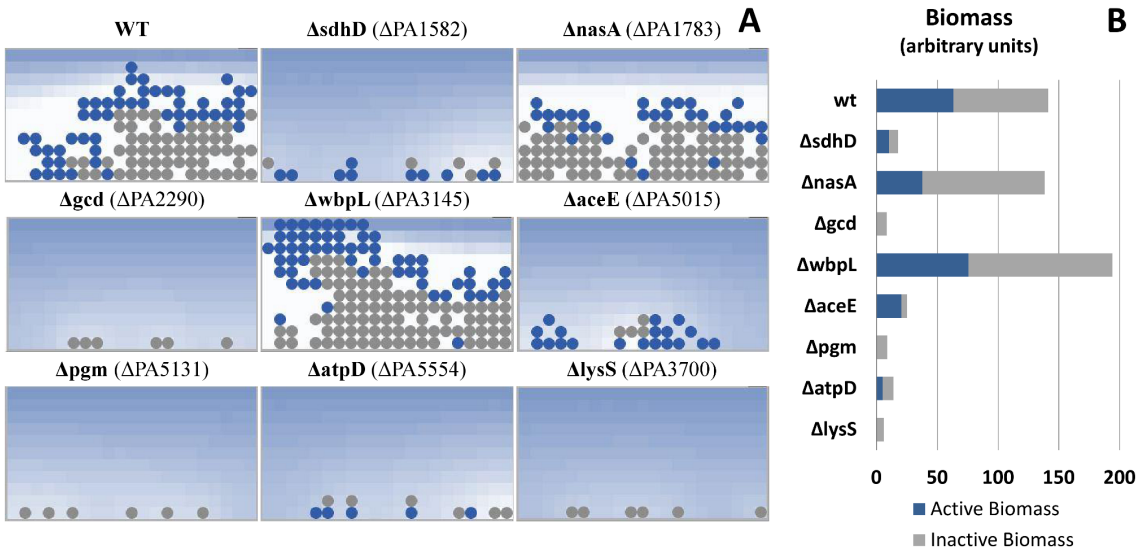


Figure 2.4: **Single-gene deletion screen.** Models of several single-deletion mutants were evaluated for biofilm formation after 200 time steps in nitrate-supplemented glucose minimal media. The wild-type (WT) model serves as a positive control. Δ*lysS* is known to be lethal, and provides a negative control. As such, the six initial cells seeded in the model never produced any additional biomass. (A) Snapshots of each multiscale simulation at time step 200. (B) Proportions of active and inactive biomass for each ABM at time step 200. Δ*sdhD*, Δ*aceE* and Δ*atpD* grew more slowly than wild-type. Δ*gcd* and Δ*pgm* appeared to have significant growth defects (final biomass only slightly more than that initially seeded). This screen is an example of a powerful analysis that is enabled by the multiscale simulations integrating spatial modeling with NetLogo and the metabolic network analysis performed in MATLAB.

an interrogation of the transition from a planktonic to biofilm state. The current work demonstrates that even simplified multiscale models can capture important biological behaviors that would be difficult or impossible to predict otherwise, and that our tool enables powerful cross-platform modeling that could be of value in multiple biomedical and other applications.

2.7 Acknowledgments

We thank our colleagues Joanna Goldberg, Shayn Peirce-Cottler, John Varga, Jennifer Bartell, and Phillip Yen for their helpful suggestions during the writing of this manuscript.

2.8 References

[1] Walpole J, Papin JA, and Peirce SM. "Multiscale Computational Models of Complex Biological Systems." In: *Annual review of biomedical engineering* 15.April (Apr. 2013), pp. 137–154. DOI: 10.1146/annurev-bioeng-071811-150104.

[2] Hayenga HN, Thorne BC, Peirce SM, and Humphrey JD. "Ensuring Congruency in Multiscale Modeling: Towards Linking Agent Based and Continuum Biomechanical Models of Arterial Adaptation". In: *Annals of Biomedical Engineering* 39.11 (2012), pp. 2669–2682. DOI: 10.1039/b000000x/This.

[3] Thorne BC, Hayenga HN, Humphrey JD, and Peirce SM. "Toward a multi-scale computational model of arterial adaptation in hypertension: verification of a multi-cell agent based model." In: *Frontiers in physiology* 2.May (Jan. 2011), p. 20. DOI: 10.3389/fphys.2011.00020.

[4] R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria, 2016.

[5] Wilensky U. *NetLogo*. Evanston, IL., 1999.

[6] MathWorks. *MATLAB and Statistics Toolbox*. Natick, Massachusetts, USA, 2012.

[7] Thiele JC and Grimm V. "NetLogo meets R: Linking agent-based models with a toolbox for their analysis". In: *Environmental Modelling & Software* 25.8 (Aug. 2010), pp. 972–974. DOI: 10.1016/j.envsoft.2010.02.008.

[8] Bengtsson H. *R.matlab - Local and remote MATLAB connectivity in R*. Lund, Sweden, 2005.

[9] Robertson SH et al. "Multiscale computational analysis of Xenopus laevis morphogenesis reveals key insights of systems-level behavior." In: *BMC systems biology* 1 (Jan. 2007), p. 46. DOI: 10.1186/1752-0509-1-46.

[10] Neidig A et al. "TypA is involved in virulence, antimicrobial resistance and biofilm formation in Pseudomonas aeruginosa." In: *BMC microbiology* 13 (Jan. 2013), p. 77. DOI: 10.1186/1471-2180-13-77.

[11] Bjarnsholt T et al. "Pseudomonas aeruginosa biofilms in the respiratory tract of cystic fibrosis patients." In: *Pediatric pulmonology* 44.6 (June 2009), pp. 547–58. DOI: 10.1002/ppul.21011.

[12] Pizarro G, Griffeath D, and Noguera D. "Quantitative cellular automaton model for biofilms". In: *Journal of Environmental Engineering* 127 (2001), pp. 782–789.

[13] Noguera D, Pizarro G, and JM R. "Modeling Biofilms". In: *Microbial Biofilms*. Ed. by Ghannoum M and O'Toole G. 2004.

[14] Oberhardt MA et al. "Genome-scale metabolic network analysis of the opportunistic pathogen Pseudomonas aeruginosa PAO1." In: *Journal of bacteriology* 190.8 (Apr. 2008), pp. 2790–803. DOI: 10.1128/JB.01583-07.

[15] Railsback SF and Grimm V. *Agent-Based and Individual-Based Modeling: A Practical Introduction*. Princeton and Oxford: Princeton University Press, 2012.

[16] Oberhardt MA, Goldberg JB, Hogardt M, and Papin JA. "Metabolic Network Analysis of Pseudomonas aeruginosa during Chronic Cystic Fibrosis Lung Infection". In: *Journal of Bacteriology* 192.20 (Oct. 2010), pp. 5534–5548. DOI: 10.1128/JB.00900-10.

[17] Stewart PS. "Diffusion in Biofilms GUEST COMMENTARIES Diffusion in Biofilms". In: 185.5 (2003). DOI: 10.1128/JB.185.5.1485.

[18] Becker SA et al. "Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox." In: *Nature protocols* 2.3 (Jan. 2007), pp. 727–38. DOI: 10.1038/nprot.2007.99.

[19] Gurobi. *Gurobi Optimizer*. Houston, TX, 2013.

[20] Gianchandani EP, Chavali AK, and Papin JA. "The application of flux balance analysis in systems biology". In: *Wiley interdisciplinary reviews. Systems biology and medicine* 2.3 (2010), pp. 372–382. DOI: 10.1002/wsbm.60.

[21] Mahadevan R, Edwards JS, and Doyle FJ. "Dynamic flux balance analysis of diauxic growth in Escherichia coli." In: *Biophysical journal* 83.3 (2002), pp. 1331–1340. DOI: 10.1016/S0006-3495(02)73903-9.

[22] Xu KD et al. "Spatial physiological heterogeneity in Pseudomonas aeruginosa biofilm is determined by oxygen availability." In: *Applied and environmental microbiology* 64.10 (Oct. 1998), pp. 4035–9.

[23] Chave J. "The problem of pattern and scale in ecology: what have we learned in 20 years?" In: *Ecology Letters* 16 (May 2013). Ed. by Bascompte J, pp. 4–16. DOI: 10.1111/ele.12048.

[24] Magliocca NR and Ellis EC. "Using Pattern-oriented Modeling (POM) to Cope with Uncertainty in Multiscale Agent-based Models of Land Change". In: *Transactions in GIS* 17.6 (Dec. 2013), pp. 883–900. DOI: 10.1111/tgis.12012.

[25] Jia T et al. "An empirical study on human mobility and its agent-based modeling". In: *Journal of Statistical Mechanics: Theory and Experiment* 2012.11 (Nov. 2012), P11024. DOI: 10.1088/1742-5468/2012/11/P11024.

[26] Trunk K et al. "Anaerobic adaptation in Pseudomonas aeruginosa: definition of the Anr and Dnr regulons." In: *Environmental microbiology* 12.6 (June 2010), pp. 1719–33. DOI: 10.1111/j.1462-2920.2010.02252.x.

[27] Borriello G, Werner E, and Roe F. "Oxygen limitation contributes to antibiotic tolerance of Pseudomonas aeruginosa in biofilms". In: *Antimicrobial agents and Chemotherapy* 48.7 (2004), pp. 2659–2664. DOI: 10.1128/AAC.48.7.2659.

[28] Manos J et al. "Gene expression characteristics of a cystic fibrosis epidemic strain of Pseudomonas aeruginosa during biofilm and planktonic growth." In: *FEMS microbiology letters* 292.1 (Mar. 2009), pp. 107–14. DOI: 10.1111/j.1574-6968.2008.01472.x.

[29] Werra P de, Péchy-Tarr M, Keel C, and Maurhofer M. "Role of gluconic acid production in the regulation of biocontrol traits of Pseudomonas fluorescens CHA0." In: *Applied and environmental microbiology* 75.12 (June 2009), pp. 4162–74. DOI: 10.1128/AEM.00295-09.

[30] Lardon La et al. "iDynoMiCS: next-generation individual-based modelling of biofilms." In: *Environmental microbiology* 13.9 (Sept. 2011), pp. 2416–34. DOI: 10.1111/j.1462-2920.2011.02414.x.

[31] Merkey BV et al. "Growth dependence of conjugation explains limited plasmid invasion in biofilms: an individual-based modelling study." In: *Environmental microbiology* 13.9 (Oct. 2011), pp. 2435–52. DOI: 10.1111/j.1462-2920.2011.02535.x.

[32] Mabrouk N, Mathias JD, and Deffuant G. "Viability and Resilience of Complex Systems". In: Understanding Complex Systems (2011). Ed. by Deffuant G and Gilbert N. DOI: 10.1007/978-3-642-20423-4.

[33] Mabrouk N, Deffuant G, Tolker-Nielsen T, and Lobry C. "Bacteria can form interconnected microcolonies when a self-excreted product reduces their surface motility: evidence from individual-based model simulations." In: *Theory in biosciences = Theorie in den Biowissenschaften* 129.1 (June 2010), pp. 1–13. DOI: 10.1007/s12064-009-0078-8.

Chapter 3

Inference of Network Dynamics and Metabolic Interactions in the Gut Microbiome

The text for this chapter has been previously published as a research article here:

Steinway SN*, Biggs MB*, Loughran TP Jr., Papin JA$^{\delta}$, Albert R$^{\delta}$. (2015). Inference of Network Dynamics and Metabolic Interactions in the Gut Microbiome. *PLOS Computational Biology*,11(6):e1004338.

* These individuals contributed equally to this work.

δ Co-corresponding authors.

3.1 Context

Steve Steinway was an MD/PhD student completing his research in Dr. Tom Loughran's lab in the Department of Medicine. Steve moved to UVA partway through his PhD program from Pennsylvania State University where he was co-mentored by Dr. Reka Albert, a computational biologist. His expertise was in building dynamic Boolean network models of protein signalling networks. He and I met at a mixer for the Jefferson Trust Big Data Fellowship. He and I shared interests in network modeling and the gut microbiome. We teamed up, and were successful in winning a Jefferson Trust Big Data Fellowship which totaled more that $44,000. Once again, we set a goal to publish our work, which required a lot of re-starting from scratch over the course of a year as project ideas failed. Finally, we hit upon a great time-series data set, an interesting approach to analyzing it and some encouraging wet lab results which supported our conclusions. I was lucky to find someone as talented and fun to work with as Steve, and we both were lucky to have supportive mentors through the process of completing this independent and seeming "side project". At this writing, Steve was completing his MD at the Pennsylvania State University School of Medicine.

3.2 Synopsis

We present a novel methodology to construct a Boolean dynamic model from time series metagenomic information and integrate this modeling with genome-scale metabolic network reconstructions to identify metabolic underpinnings for microbial interactions. We apply this in the context of a critical health issue: clindamycin antibiotic treatment and opportunistic *Clostridium difficile* infection. Our model recapitulates known dynamics of clindamycin antibiotic treatment and *C. difficile* infection and predicts therapeutic probiotic interventions to suppress *C. difficile* infection. Genome-scale metabolic network reconstructions reveal metabolic differences between community members and are used to explore the role of metabolism in the observed microbial interactions. *In vitro* experimental data validate a key result of our computational model, that *B. intestinihominis* can in fact slow *C. difficile* growth.

3.3 Introduction

Human health is inseparably connected to the billions of microbes that live in and on us. Current research shows that our associations with microbes are, more often than not, essential for our health [1]. The microbes that live in and on us (collectively our "microbiome") help us to digest our food, train our immune systems, and protect us from pathogens [2, 3]. The gut microbiome is an enormous community, consisting of hundreds of species and trillions of individual interacting bacteria [4]. Microbial community composition often persists for years without significant change [5].

When change comes, however, it can have unpredictable and sometimes fatal consequences. Acute and recurring infections by *Clostridium difficile* have been strongly linked to changes in gut microbiota [6]. The generally accepted paradigm is that antibiotic treatment (or some other perturbation) significantly disrupts the microbial community structure in the gut, which creates a void that *C. difficile* will subsequently fill [7–10]. Such infections occur in roughly 600,000 people in the United States each year (this number is on the rise), with an associated mortality rate of 2.3% [11]. Each year, health care costs associated with *C. difficile* infection are in excess of $3.2 billion [11]. An altered gut flora has further

been identified as a causal factor in obesity, diabetes, some cancers and behavioral disorders [12–17].

What promotes the stability of a microbial community, or causes its collapse, is poorly understood. Until we know what promotes stability, we cannot design targeted treatments that prevent microbiome disruption, nor can we rebuild a disrupted microbiome. Studying the system level properties and dynamics of a large community is impossible using traditional microbiology approaches. However, network science is an emerging field which provides a powerful framework for the study of complex systems like the gut microbiome [18–23]. Previous efforts to capture the essential dynamics of the gut have made heavy use of ordinary differential equation (ODE) models [24, 25]. Such models require the estimation of many parameters. With so many degrees of freedom, it is possible to overfit the underlying data, and it is difficult to scale up to larger communities [26, 27]. Boolean dynamic models, conversely, require far less parameterization. Such models capture the essential dynamics of a system, and scale to larger systems. Boolean models have been successfully applied at the molecular [28, 29], cellular [20], and community levels [30]. Here we present the first Boolean dynamic model constructed from metagenomic sequence information and the first application of Boolean modeling to microbial community analysis.

We analyze the dynamic nature of the gut microbiome, focusing on the effect of clindamycin antibiotic treatment and *C. difficile* infection on gut microbial community structure. We generate a microbial interaction network and dynamical model based on time-series data from metagenome data from a population of mice. We present the results of a dynamic network analysis, including steady-state conditions, how those steady states are reached and maintained, how they relate to the health or disease status of the mice, and how targeted changes in the network can transition the community from a disease state to a healthy state. Furthermore, knowing how microbes positively or negatively impact each other—particularly for key microbes in the community—increases the therapeutic utility of the inferred interaction network. We produced genome-scale metabolic reconstructions of the taxa represented in this community [31], and probe how metabolism could—and could not—contribute to the mechanistic underpinnings of the observed interactions. We present validating experimental evidence consistent with our computational results, indicating that a member of the normal gut flora, *Barnesiella*, can in fact slow *C. difficile* growth.

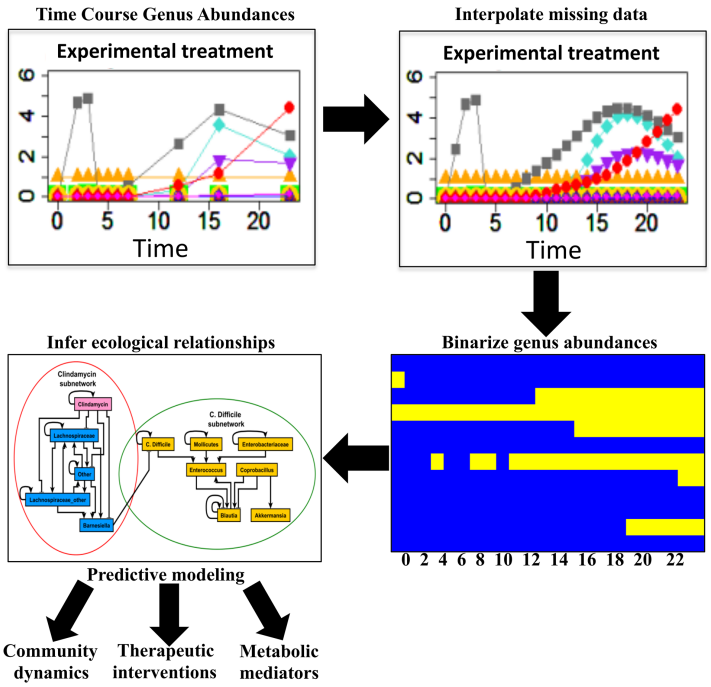


Figure 3.1: **Dynamic analysis workflow.** Time course genus abundance information was acquired from metagenomic sequencing of mouse gastrointestinal tracts under varying experimental conditions. Missing time points from experimental data were estimated such that genus abundances existed at the same time points across all treatment groups. Next, genus abundances were binarized such that Boolean regulatory relationships could be inferred. A dynamic Boolean model was constructed to explore gut microbial dynamics, therapeutic interventions, and metabolic mediators of bacterial regulatory relationships.

3.4 Methods

Data Sources

Buffie *et al.* reported treating mice with clindamycin and tracking microbial abundance by 16S sequencing [32]. Mice treated with clindamycin were more susceptible to *C. difficile* infection than controls. The collection of 16S sequences corresponding to these experiments was analyzed by Stein *et al.* [24]. First, Stein *et al.* aggregated the data by quantifying microbial abundance at the genus level. Abundances of the ten most abundant genera and an other group were presented as operational taxonomic unit (OTU) counts per sample. We use the aggregated abundances from Stein *et al.* as the starting point for our modeling pipeline (Figure 3.1).

This processed dataset consisted of nine samples and three treatment groups (n = 3 replicates per treatment group). The first treatment group (here called "Healthy") received spores of *C. difficile* at t = 0 days,

and was used to determine the susceptibility of the native microbiota to invasion. The second treatment group (here called "clindamycin treated") received a single dose of clindamycin at t = -1 days to assess the effect of the antibiotic alone, and the third treatment group (here called clindamycin+ C. difficile treated) received a single dose of clindamycin (at t = -1 days) and, on the following day, was inoculated with *C. difficile* spores (Figure S1A). Under the clindamycin+ *C. difficile* treatment group conditions, *C. difficile* could colonize the mice and produce colitis; however this was not possible under the first two treatment group conditions.

Interpolation of Missing Genus Abundance Information

The gut bacterial genus abundance dataset included some variation in terms of time points in which genera were sampled. That is, genus abundances were measured between 0 to 23 days; however, not all samples had measurements at all the time points (Figure S1A). Particularly, the healthy population only included time points at 0, 2, 6, and 13 days and Sample 1 of clindamycin+ *C. difficile* treated population was missing the 9 day time point. Missing abundance values for these 4 points were estimated using an interpolation approach (Figure S1B). For healthy samples, the 16 and 23 day time points could not be interpolated as the last experimentally identified time point for these samples is at 13 days. The assumption of the approximated polynomial for these samples is that extrapolated data points are linear using the slope of the interpolating curve at the nearest data point. Because genera abundances are fairly stable across time in this treatment group (i.e. the slope of most of the genera abundances is approximately zero), extrapolating two time points was deemed reasonable. A principal component analysis was completed on the interpolated data (Figure 3.2A) and shows that the interpolated time series bacterial genus abundance data clusters by experimental treatment group in the first two principal components. Furthermore, the results of the binarization for the healthy population suggest that interpolation did not have any concerning effects on the 16 and 23 day time points (Figure S2).

Natural cubic spline interpolation was used to estimate genus abundances at missing time points in some samples. A cubic spline is constructed of piecewise third order (cubic) polynomials which pass through the known data points and has continuous first and second derivatives across all points in the dataset. Natural cubic spline is a cubic spline that has a second derivative equal to zero at the end points of the dataset [33]. Natural splines were interpolated such that all datasets had time points at single day intervals through the 23 day time point (Figure S1B).

Network Modeling Framework

We use a Boolean framework in which each network node is described by one of two qualitative states: ON or OFF. We chose this framework because of its computational feasibility and capacity to be constructed with minimal and qualitative biological data [34]. The ON (logical 1) state means an above threshold abundance of a bacterial genus whereas the OFF (logical 0) state means below-threshold genus absence. The putative biological relationships among genera are expressed as mathematical equations using Boolean operators [29, 34]. We inferred putative Boolean regulatory functions for each node, which are able to best capture the trends in the bacterial abundances. These rules, (edges in the interaction network) can be assigned a direction, representing information flow, i.e. effect from the source (upstream) node to the target (downstream) node. Furthermore, edges can be characterized as positive (growth promoting) or negative (growth suppressing). An additional layer of network analysis is the dynamic model, which is used to express the behavior of a system over time by characterizing each node by a state variable (e.g., abundance) and a function that describes its regulation. Dynamic models can be categorized as continuous or discrete, according to the type of node state variable used. Continuous models use a set of differential equations; however, the paucity of known kinetic details for inter-genus and/or inter-species interactions makes these models difficult to implement.

Binarization

Genus abundance data was binarized (converted to a presence-absence dataset) to enable inference of Boolean relationships for modeling applications. We adapted a previously developed approach called iterative k-means binarization with a clustering depth of 3 (KM3) for this purpose [35]. This approach was employed because binarized data is able to maintain complex oscillatory behavior in Boolean models constructed from this data, whereas other binarization approaches fail to maintain these features [35].

Briefly, this approach uses k-means clustering with a depth of clustering d and an initial number of clusters $k = 2^d$. In each iteration, data for a specific genus G are clustered into k unique clusters $C_G^1, , C_G^k$, then for each cluster, C_G^n, all the values are replaced by the mean value of C_G^n. For the next iteration, the value of d is decreased and clustering is repeated. This methodology is repeated until $d = 1$. This approach, with $d = 3$ (referred to here

as KM3 binarization) has previously been demonstrated as a superior binarization methodology to other binarization approaches for Boolean model construction because it conserves oscillatory behavior [35]. These analyses were performed using custom Python code based on a previously written algorithm [35] and is available in the supplemental materials.

Because KM3 binarization has a stochastic component (the initial grouping of binarization clusters), we employed KM3 binarization on the entire bacterial genus abundance time series dataset 1000 times. The average binarization for each sample (Figure S2) was used to determine the most probable binarized state of each genus in each sample at each time point (Figure S3). A principal component analysis of the most probable binarized genus abundances for each sample demonstrates that as with the continuous time series abundances (Figure 3.2A), binarized bacterial genus abundance data cluster by experimental treatment group (Figure 3.2B). For inference of Boolean rules from the binarized genus abundances (Figure S3), the consensus of two of three samples for each treatment population was used as the binarized state of each genus at each time point in each sample (Figure 3.2C).

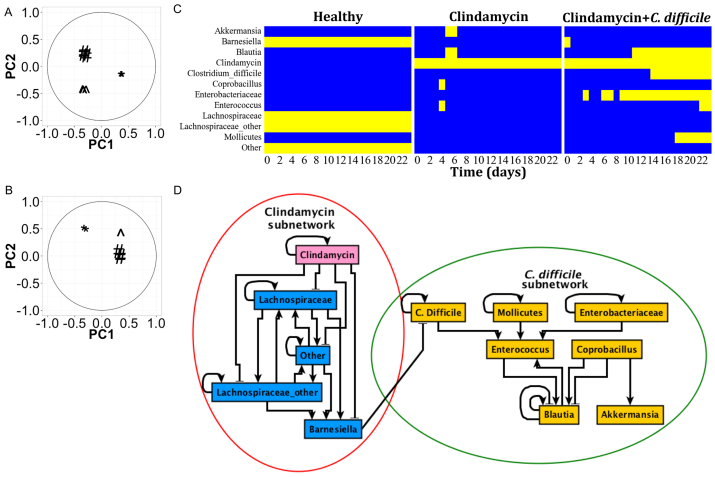


Figure 3.2: **Construction of a network model of the gut microbiome from time course metagenomic genus abundance information.** Principal component analysis coefficients associated with each sample in the metagenomic genus abundance dataset was completed for A) interpolated genus abundances and B) binarized interpolated genus abundances. "*" = Healthy; "∧" = clindamycin treated; "#" = clindamycin+ *C. difficile* treated. C) Consensus binarization of genus abundance information. Each heatmap represents the consensus binarization for each treatment group. The horizontal axis represents the day of the experiment that the sample came from. The vertical axis represents the specific genera being modeled. Each genus was binarized to a 1 (ON; above activity threshold) or 0 (OFF; below activity threshold). D) Interaction rules were inferred from the binarized data. The interaction rules were simplified for visualization (compound rules were broken into simple one-to-one edges).

Inference of Boolean Rules from Time Series Genus Abundance Information

The Best-fit extension was applied to learn Boolean rules from the binarized time series genus abundance information [36]. For each variable (genus) X_i in the binarized time series genus abundance data, Best-fit identifies the set of Boolean rules with k variables (regulators) that explains the variable's time pattern with the least error size. The algorithm uses partially defined Boolean functions $pdBf(T, F)$, where the set of true (T) and false vectors (F) are defined as $T = \{X' \in \{0,1\}^k : X_i(t+1) = 1\}$ and $F = \{X' \in \{0,1\}^k : X_i(t+1) = 0\}$. Intuitively, the partial Boolean function summarizes the states of the putative regulators that correspond to a turning ON (T) or turning OFF (F) of the target variable. The error size ϵ of $pdBf(T, F)$ is defined as the minimum number of inconsistencies within X' that best classifies the T and F values of the dataset. The Best-Fit extension works by identifying smallest size X' for X_i. For more detailed information refer to [36]. In line with this, we considered the most parsimonious representation of the rules with the smallest ϵ. If the most parsimonious rule was self-regulation, we also considered rules with the same ϵ that included another regulator. If multiple rules fit these criteria for a given X_i, it implied that they can independently represent the inferred regulatory relationships. In cases where the alternatives had the same value of (non-zero) ϵ, we explored combinations (such as appending them by an OR rule) and used the combination that best described the experimentally observed final (steady state) outcomes. For example, we combined the two alternative rules for *Blautia* with an OR relationship. In the case of *Barnesiella*, we chained three rules ("Other", "Lachnospiraceae_other","Lachnospiraceae") by an OR relationship, and "not Clindamycin" by an AND relationship to incorporate the loss of *Barnesiella* in the presence of clindamycin (Figure 3.2C). This was also done for rules for "Lachnospiraceae", "Lachnospiraceae_other" and "Other" and all four nodes attained the same rule. There are six nodes with multiple inferred (alternative) rules: "Barnesiella", "Blautia", "Enterococcus", "Lachnospiraceae", "Lachnospiraceae_other", and "Other" had 4, 2, 5, 4, 4, and 4 rules, respectively. The six other nodes had a single inferred rule. The network in Figure 3.2C represents the union of all of the alter-

native rules produced by Best-Fit, or in other words, it is a super-network of all alternative rules. Any alternative networks would be a sub-network of what we show. A strongly connected component between the nodes inhibited by clindamycin is a feature of the vast majority of these sub-networks. We used the implementation of Best-Fit in the R package BoolNet [37].

Dynamic Analysis

Dynamic analysis is performed by applying the inferred Boolean functions in succession until a steady state is reached. Boolean models and discrete dynamic models in general focus on state transitions instead of following the system in continuous time. Thus, time is an implicit variable in these models. The network transitions from an initial condition (initial state of the bacterial community) until an attractor is reached. An attractor can be a fixed point (steady state) or a set of states that repeat indefinitely (a complex attractor). The basin of attraction refers to the set of initial conditions that lead the system to a specific attractor. For the network under consideration, the complete state space can be traversed by enumerating every possible combination of node states (212) and applying the inferred Boolean functions (or update rules) to determine paths linking those states. The state transition network describes all possible community trajectories from initial conditions to steady states, given the observed interactions between bacteria in the community.

We made use of two update schemes to simulate network dynamics: synchronous (deterministic) and asynchronous (stochastic). Synchronous models are the simplest update method: all nodes are updated at multiples of a common time step based on the previous state of the system. The synchronous model is deterministic in that the sequence of state transitions is definite for identical initial conditions of a model. In asynchronous models, the nodes are updated individually, depending on the timing information, or lack thereof, of individual biological events. In the general asynchronous model used here, a single node is randomly updated at each time step [38]. The general asynchronous model is useful when there is heterogeneity in the timing of network events but when the specific timing is unknown. Due to the heterogeneous mechanisms by which bacteria interact, we made the assumption of time heterogeneity without specifically known time relationships. Synchronous and asynchronous Boolean models have the same fixed points, because fixed points are independent of the implementation of time. However, the basin of attraction of each fixed point (i.e. the initial conditions that lead to each fixed point) may differ between synchronous and asynchronous models (Table S2). For identification of all of the fixed points in the network (the attractor landscape), the synchronous updating scheme was used. However, for the perturbation analysis, the asynchronous updating scheme was used because it more realistically models the possible trajectories in a stochastic and/or time-heterogeneous system. The simulations of the gut microbiome model were performed using custom Python code built on top of the BooleanNet Python library, which facilitates Boolean simulations [39]. Our custom Python code is available in the supplemental materials.

Perturbation Analysis

To capture the effect of removal (knockout) or addition (probiotic; forced over abundance) of genera, modification of the states/rules to describe removal or addition states were performed. These modifications were implemented in BooleanNet by setting the corresponding nodes to either OFF (removal) or ON (addition) and then removing the corresponding updating rules for these nodes for the simulations. By examining many such forced perturbations, we can identify potential therapeutic strategies, many of which may not be obvious or intuitive, particularly as network complexity increases. We used asynchronous update when simulating the effect of perturbations on the microbial communities. In each case we performed 1000 simulations and report the percentage of simulations that achieve a certain outcome.

Generating Genus-Level Genome-Scale Metabolic Reconstructions

To generate draft metabolic network reconstructions for each of the ten genera in the paper, we first obtained genome sequences for representative species by searching the Genomes database of the National Center for Biotechnology Information (NCBI). Complete genomes for the first ten (or if less than ten, all) species within the appropriate genus were downloaded. During the process of reconstructing genus-level metabolic reconstructions, some genera were underrepresented (fewer than 10 species genomes) in the NCBI Genome database, including *Akkermansia*, *Barnesiella* and *Coprobacillus* (Table S3). The search result order is based on record update time, and so it is quasi-random. Genomes were uploaded to the rapid annotations using subsystems technology (RAST) server for annotation [40]. Draft metabolic network reconstructions were generated by providing the RAST annotations to the Model SEED service [41]. Metabolic network reconstructions were downloaded in .xls format. Genus-level metabolic reconstructions were

produced by taking the union of all species-level reconstructions corresponding to each genus, as has been done previously [42]. The one exception was *C. difficile*, which was produced by taking the union of three strain-level reconstructions.

Subsystem Enrichment Analysis

Subsystems were defined as the Kyoto Encyclopedia of Genes and Genomes (KEGG) map with which each reaction was associated [43, 44]. These associations were determined based on annotations in the Model SEED database [41]. To quantify enrichment, the complete set of unique reactions from all genus-level reconstructions was pooled, and the subsystem annotations corresponding to those reactions were counted. To determine enrichment for a given subset of the community (either a single genus-level reconstruction, or a set of reconstructions corresponding to a subnetwork), the subsystem occurrences were counted within the subset. The probability of a reconstruction containing N total subsystem annotations, with M or more occurrences of subsystem I, was determined by taking the sum of a hypergeometric probability distribution function (PDF) from M to the total occurrences of subsystem I in the overall population. Enrichment analysis was performed in Matlab [45].

Identifying Seed Sets and Defining Metabolic Competition and Mutualism Scores

To quantify metabolic interactions, we started by utilizing the seed set detection algorithm developed by Borenstein *et al.* [46, 47]. The algorithm follows three steps:

1. The genome-scale metabolic network reconstruction is reduced into simple one-to-one edges, such that for each reaction, each substrate and product pair forms an edge (e.g. $A + B \rightarrow C$ would become $A \rightarrow C$ and $B \rightarrow C$).

2. The network is divided into strongly connected components, those groups of nodes for which two paths of opposite directions (e.g. $A \rightarrow B$ and $B \rightarrow A$) exist between any two nodes in the group.

3. Nodes (and strongly connected components with five or fewer nodes) for which there are exclusively outgoing edges are defined as "inputs" to the model, or seed metabolites.

The rationale is that metabolites that feed into the network, but cannot be produced by any reactions within the network, must be obtained from the environment.

Competition metrics were generated following the process of Levy and Borenstein [46]. For a given pair of genera, the competition score is defined as:

$$CompScore_{ij} = \frac{|SeedSet_i \cap SeedSet_j|}{|SeedSet_i|}$$

Here $SeedSet_i$ is the set of obligatory input metabolites to the metabolic network reconstruction for genus i, and $|SeedSet_i|$ is the number of metabolites contained in $SeedSet_i$. The competition score indicates the fractional overlap of inputs that genus i shares with genus j, and so ranges between zero and one. The higher the score, the more similar the metabolic inputs to the two networks, making competition more likely.

For a given pair of genera, the mutualism score is defined as:

$$MutualismScore_{ij} = \frac{|SeedSet_i \cap \neg SeedSet_j|}{|SeedSet_i|}$$

Here $\neg SeedSet_j$ is the set of metabolites that can be produced by the metabolic network for species j (i.e. all non-seed metabolites). The mutualism score indicates the fractional overlap of inputs that genus i consumes which genus j can potentially provide. The mutualism score ranges between zero and one. The higher the score, the more potential there is for nutrient sharing between species. While the score does not measure mutualism per se (it cannot necessarily distinguish between other interactions such as commensalism or amenalism [48]), for simplicity, we will refer to these scores as the competition and mutualism scores.

All metabolic reconstructions, seed sets, competition scores and mutualism scores are available in the supplemental materials. Seed set generation was performed using custom Matlab scripts, which are available in the supplement. Statistical tests were performed in R [49].

Co-culture and Spent Media Experiments

Barnesiella intestinihominis DSM 21032 and *Clostridium difficile* VPI 10463 were grown anaerobically in PRAS chopped meat medium (CMB) (Anaerobe Systems, Morgan Hill, CA) at 37 °C. To prepare *B. intestinihominis* spent medium, *B. intestinihominis* was grown in CMB until stationary phase (44 hours). The saturated culture was centrifuged, and the supernatant was filter sterilized (0.22 μM pore size). Growth curves were obtained by inoculating batch cultures in 96-well plates and gathering optical density measurements (870 nm) using a small plate reader that fits in the anaerobic chamber [50]. Single cultures were inoculated from overnight liquid culture to a starting density of 0.01. The co-cultures were started at a 1:1 ratio, for a total

starting density of 0.02. Optical density was measured every 2 minutes for 24 hours, and the resulting growth curves were analyzed in Matlab [45]. Maximum growth rates were calculated by fitting a smooth line to each growth curve, and finding the maximum growth rate from among the instantaneous growth rates over the whole time course: $[log(OD_{t+1}) - log(OD_t)]/[t_{+1} - t]$. The achieved bacterial density—area under the growth curve (AUC)—in a culture was calculated by integrating over the growth curve in each experiment using the "trapz()"function in Matlab. It can be thought of as representing the total biomass produced over time. The simply additive null model was calculated by fitting a Lotka-Volterra model [24] to the single cultures for both *B. intestihominis* and *C. difficile*. The null model of co-culture (assuming zero interaction between species) was simulated by using the parameters from single culture, and summing the predicted OD870 values.

All scripts used to analyze the data are available at: bitbucket.org/gutmicrobiomepaper/.

3.5 Results

Processing of a Microbial Genus Abundance Dataset for Network Inference

To capture the dynamics of inter-genus interactions in the intestinal tract we employed a pipeline (Figure 3.1) which translates metagenomic genus abundance information into a dynamic Boolean model. This approach involves three steps: 1) discretization (binarization) of genus abundances, 2) learning Boolean relationships among genera, and 3) translation of genus associations into a Boolean (discrete) dynamic model.

Construction of a Dynamic Network Model from Binarized Time Series Microbial Genus Abundance Information

Boolean rules (Table S1e) were inferred from the time series binarized genus abundances using an implementation of the Best-fit extension [36] in the R Boolean network inference package BoolNet (see Methods) [37]. A network of 12 nodes and 33 edges was inferred (Figure 3.2D). The inferred interaction network has a clustered structure: the cluster (subnetwork) containing the two *Lachnospiraceae* nodes and *Barnesiella* is strongly influenced by clindamycin whereas the other subnetwork is largely independent of the first, except for the single edge between *Barnesiella* and *C. difficile* (Fig 3.2D). In fact, *Lachnospiraceae* nodes, *Barnesiella* and the group of "Other" genera form a strongly connected component; that is, every node is reachable from every other node. Most nodes of the second subnetwork are positively influenced by *C. difficile*, with the exception of *Coprobacillus*, for which no regulation by other nodes was inferred, and *Akkermansia*, which is inferred to be regulated only by *Coprobacillus*. These latter two genera are transiently present (around day 5) in the clindamycin treatment group, but they do not appear in the final states of any of the treatment groups (see Figure S1). This network structure is consistent with published data in which the dominant Firmicutes (*Lachnospiraceae*) and Bacteroidetes (*Barnesiella*) are devastated by antibiotic administration [51, 52]. Furthermore, the clustered structure (Figure 3.2D) supports the established mechanism of *C. difficile* colitis: loss of normal gut flora, which normally suppresses opportunistic infection (clindamycin cluster), and the presence of *C. difficile* at a minimum inoculum (*C. difficile* cluster) [10, 53]. The network clusters have a single route of interaction between *Barnesiella* and *C. difficile*.

The negative influence of *Barnesiella* on *C. difficile* is in agreement with recently published findings in which *Barnesiella* was strongly correlated with *C. difficile* clearance [54]. The role of *Barnesiella* as an inhibitor of another pathogen (vancomycin-resistant Enterococci (VRE)) has been shown in mice [55], which is also visible in the network model as an indirect relationship between *Barnesiella* and *Enterococcus* (Figure 3.2D). Related species of Bacteroidetes have been shown to play vital roles in protection from *C. difficile* infection in mice [56][56]. Furthermore, the network structure shows that *Lachnospiraceae* positively interacts with *Barnesiella*, leading to an indirect suppression of *C. difficile*. Interestingly, the two *Lachnospiraceae* nodes and the "Other" node form a strongly connected component, suggesting a similar role in the network, particularly in promoting growth of *Barnesiella*, which directly suppresses *C. difficile*. In support of this finding, *Lachnospiraceae* has been shown to protect mice against *C. difficile* colonization [52, 57]. Therefore, the structure of the network is both a parsimonious representation of the current data set, and is supported by literature evidence.

We applied dynamic analysis using the synchronous updating scheme (see Methods) to determine all the possible steady states of the microbiome network model. In a 12 node network, there are 212 possible network states. We employed model simulations using the synchronous updating scheme to visit all possible network states and identify all fixed points of the model. Exploration of the steady states of this network reveals 23 possible fixed point attractors (Figure S4). Three of the identified attractors (Figure 3.3A) are in exact agreement with the experimentally identified terminal time points of bina-

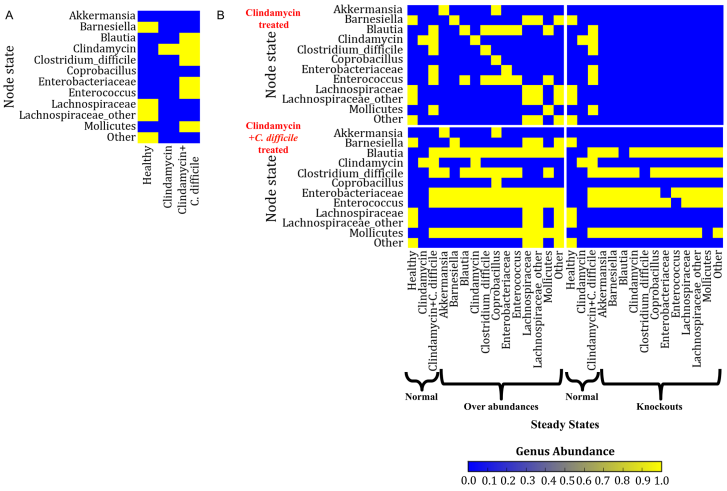


Figure 3.3: **Steady states and node perturbations in the gut microbiome model.** A) Heatmap of the three steady states in the gut microbiome model. These steady states are identical to steady states identified in the three experimental groups. B) The effect of node perturbations represented by four heatmaps. On the y-axis of each of the four heatmaps are nodes (genera) in each steady state. On the x-axis of each of the four heatmaps are the steady states found under normal model conditions (i.e. no node perturbations) and also the specific perturbation of a single network node. The two heatmaps in the left column of the figure demonstrate the effect of addition (forced overabundance) of individual genera, and the two heatmaps in the right column of the figure demonstrate the effect of removal (knockout) of individual genera. The top row heatmaps show the effect of node perturbations on the clindamycin treated group and the bottom row heatmaps show the effect of node perturbations on the clindamycin+ *C. difficile* treatment group. * Genus abundance of 0 means present in 0% of asynchronous simulations and is indicated in blue; Genus abundance of 1 means present in all (100%) of asynchronous simulations, shown in yellow. n = 1000 simulations were applied for all Boolean model simulations.

rized genus abundances (Figure 3.2C). These attractors make up a small subset of the entire microbiome network state space (Table S2).

The attractor landscape can be divided into six groups based on abundance patterns they share (Figure S4). Group 1 is made up of a single attractor wherein all genera are absent (OFF). The second group attractor consists of the experimentally defined healthy state (Attractor 2) and genera in the *C. difficile* subnetwork which can be abundant (ON) independent of the clindamycin subnetwork. The third grouping has the clindamycin treated steady state (Attractor 7) and genera in the *C. difficile* subnetwork that can survive in the presence of the clindamycin. Group 4 contains the clindamycin+*C. difficile* steady state (Attractor 12) and its subsets in which one or both of the source nodes *Mollicutes* and *Enterobacteriaceae* are absent. Group 5 contains attractors in which clindamycin is absent and *C. difficile* is present. Even if clindamycin is absent, our model suggests that *C. difficile* can thrive if *Lachnospiraceae* and *Barnesiella* are absent, i.e. these states represent a clindamycin-independent loss of *Lachnospiraceae* and *Barnesiella*. Lastly, group 6 attractors have both clindamycin and *C. difficile* as OFF. *Blautia* and *Enterococcus* are always abundant in these attractors. Indeed, because of the mutual activation between *Blautia* and *Enterococcus* they always appear together. Attractors in this group may also include the abundance (ON state) of the source nodes *Mollicutes* and *Enterobacteriaceae*.

Perturbation Analysis

We next explored the perturbation of genera in the gut microbiome network model. We considered the clinically relevant question of which perturbations might alter the microbiome steady states produced by clindamycin or clindamycin+*C. difficile* treatment after clindamycin treatment was removed. Thus, we considered the clindamycin-treated steady state (Attractor 7 in Figure S3) and the clindamycin+*C. difficile* treated steady state (Attractor 12) as initial conditions and assumed that clindamycin treatment was stopped. Our simulations, employing asynchronous update (see Methods), indicate that for both initial conditions, only the state of clindamycin changes after the treatment is stopped; these steady states become Attractor 1 and Attractor 19, respectively (S4 Fig). In other words, the steady states remain identical in the absence of clindamycin. We next explored the effect of addition (overabundance; Figure 3.3B, left column) and removal (knockout; Figure 3.3B, right column) of individual genera, simultaneously with the stopping of clindamycin treatment, on the model predicted steady states. For the perturbation analysis, the model was initialized from the clindamycin treated steady state (Figure 3.3B, top row) or the clindamycin+*C. difficile* steady state (Figure 3.3B, bottom row). From the clindamycin treated state, addition of *Lachnospiraceae* or "Other" nodes restores the healthy steady state; however, no removal restore the healthy steady state (Figure 3.3B). From the clindamycin+*C. difficile* state, addition of *Barnesiella*, *Lachnospiraceae*, or "Other" nodes lead to a shift toward the healthy steady state (suppression of *C. difficile*).

Generating Genus-Level Metabolic Reconstructions

Species-level reconstructions from the genus *Enterobacteriaceae* contained the most reactions on average (1335), while those from *Mollicutes* contained the least (485) (Table S3). The *Barnesiella* and *Enterococcus* reconstructions contained the most unique reactions (Table S4) and, interestingly, also displayed more overlap in reaction content between each other (503 reactions) than was observed between any other pair of reconstructions (Table S5). *Lachnospiraceae* and *Barnesiella* had the next highest degree of overlap (424 reactions). *Mollicutes* and *Coprobacillus* had the least degree of overlap (363 reactions) (Table S5). Note that the metabolic reconstructions produced by the SEED framework are draft quality, and as such, may lack the predictive power of well-curated metabolic reconstructions.

Subsystem Enrichment Analysis

Enrichment analysis was performed for the 99 unique subsystem annotations that were observed in the community. 22 subsystems displayed interesting enrichment patterns with respect to the structure of the interaction network (Figure 3.4). The subsystems for glycolysis/gluconeogenesis and nucleotide sugars metabolism are enriched in all taxa, highlighting the fact that all taxa contain relatively full complements of reactions within those subsystems. Interestingly, *C. difficile* is highly enriched for reactions in cyanamino acid metabolism compared to all other genera. Lipopolysaccharide (LPS) biosynthesis and cyanoamino acid metabolism subsystems are differentially enriched between *C. difficile* and both *Barnesiella* and *Lachnospiraceae*. Between *Barnesiella* and *Enterococcus*, *Barnesiella* is more highly enriched for d-glutamine and d-glutamate metabolism, pantothenate and CoA biosynthesis, LPS biosynthesis. With respect to Enterococcus, Barnesiella is less highly enriched in pyrimidine metabolism, and phenylalanine, tyrosine, and tryptophan biosynthesis.

Generating Metabolic Competition and Mutualism Scores

The metabolic reconstructions were used to explore the potential metabolic underpinnings of the inferred interaction network. Competition scores were generated for all pairwise relationships between the genera considered in the model (self-edges were excluded). The two *Lachnospiraceae* genera were treated as metabolically identical, and the "Other" group was excluded. We grouped pairs of genera into five groups based on being connected

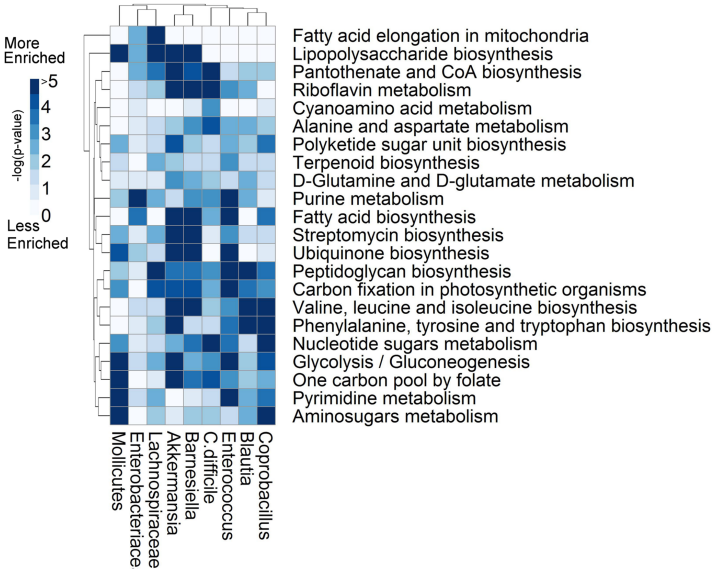


Figure 3.4: **Subsystem enrichment analysis highlights metabolic differences between taxa.** The p-values from the enrichment analysis are log-transformed and negated, such that darker regions indicate greater enrichment. The enrichment analysis quantifies the likelihood that a given subsystem (row) would be as highly abundant as observed within a given metabolic reconstruction (column) by chance alone. A subset of 22 interesting subsystems is shown here. Subsystems of interest include those for which all taxa are enriched, such as glycolysis, and nucleotide sugars metabolism, highlighting the fact that all taxa contain relatively full compliments of reactions within those subsystems. Similarly, subsystems for which a single genus differs from the remaining genera are interesting, such as cyanoamino acid metabolism, where *C. difficile* is highly enriched for reactions in that subsystem. Some subsystems are differentially enriched between *Barnesiella* and *Lachnospiraceae*, and *C. difficile* such as lipopolysaccharide biosynthesis and cyanoamino acid metabolism.

by a positive or negative edge, a negative or positive path (meaning an indirect relationship), or no path. A positive relationship was found between competition score and edge type in the interaction network (i.e. positive edges tend to have a higher competition score), which was not statistically significant, perhaps due to the small sample size (p-value = 0.058 by one-sided Wilcoxon rank sum test) (Figure S5A). The mutualism score did not display any obvious trends with respect to the network structure (Figure S5B). All pairs with inferred edges exhibited relatively high competition scores and low mutualism scores (Figure S5C). *Barnesiella*, a key inhibitor of *C. difficile* in the interaction network, holds the second smallest competition score with *C. difficile* (Figure 3.5A). *Barnesiella* and *C. difficile* also have the high-

est mutualism score among all interacting pairs in the network (Figure S5C).

The positive relationship between edge type and competition score suggests that more metabolic similarity between genera tends to foster positive interaction. The converse is also true, where less metabolic similarity tends to foster negative interactions (Figure S5A). Here, "positive/negative interaction" is derived from the Boolean model, where a positive edge between species A and B indicates that if A is ON at time t, then B is likely to turn ON at $t+1$.

Co-culture and Spent Media Experiments

Barnesiella intestinihominis was chosen as a representative species for the genus *Barnesiella* for the *in vitro* experiments. *C. difficile* grew more slowly in *B. intestinihominis* spent media (n = 16, p-value < 0.005, by one-sided Wilcoxon rank sum test) (Figure 3.5B). The co-culture with both *B. intestinihominis* and *C. difficile* grew more slowly than *C. difficile* alone (n = 16, p-value < 0.05, by one-sided Wilcoxon rank sum test) (Figure 3.5B). *C. difficile* area under the growth curve (AUC), a measure of the achieved bacterial density over the experiment, was not statistically different between growth in fresh media and *B. intestinihominis* spent media (n = 16, p-value = 0.22 by one-sided Wilcoxon rank sum test). However, the co-culture displayed a much lower AUC than expected under a null model of interaction (in which the two species do not interact) (Figure 3.5C). Examining the co-culture growth curve, it maintained a consistently lower density than a null model (Figure 3.5D).

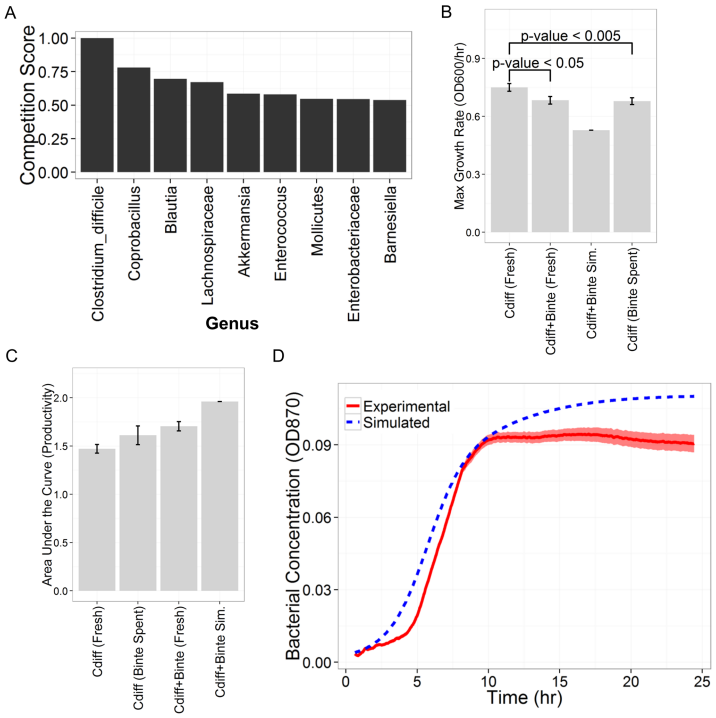


Figure 3.5: **Metabolic competition scores and *in vitro* data indicate a non-metabolic interaction mechanism.** A) Competition scores for all pairs of genera with *C. difficile*. Notice that *Barnesiella* has nearly the lowest competition score. B) Maximum growth rates for all growth conditions. *C. difficile* grew more slowly in *B. intestinihominis* spent media (n = 16, p-value $<$ 0.005, by one-sided Wilcoxon rank sum test). The co-culture with both *B. intestinihominis* and *C. difficile* grew more slowly than *C. difficile* alone (n = 16, p-value < 0.05, by one-sided Wilcoxon rank sum test). C) Area under the curve (AUC) was not significantly different for *C. difficile* in fresh media or *B. intestinihominis* spent media (n = 16, p-value = 0.22 by one-sided Wilcoxon rank sum test). D) The experimental (red, solid line) and simulated (blue, dashed line) co-culture growth curves. "Binte" indicates *B. intestinihominis*, while "Cdiff" stands for *C. difficile*. On average, the experimental co-culture growth curves maintained a lower density than the simply additive null model. Error bars represent the standard error of the mean from 16 independent replicates.

3.6 Discussion

Here we have developed a novel strategy for generating a dynamic model of gut microbiota composition by inferring relationships from time series metagenomic data (Figure 3.1). To our knowledge, this is the first Boolean dynamic model of a microbial interaction network and the first Boolean model inferred from metagenomic sequence information. Metagenomic sequencing is a powerful tool that tells us the consequences of microbial interaction—changes in bacterial abundance. Bacterial interactions are, in fact, mediated by the many chemicals and metabolites the bacteria use and produce. In a network sense these relationships are a bipartite graph; bacterial genera produce chemicals/metabolites, which have an effect on other bacteria. Because there is no comprehensive source for the bacterial metabolites and their effect on other bacterial genera, we infer the effects of genera on each other from the relative abundances of genera in a set of microbiome samples, and we employ genome-scale metabolic reconstructions to gain insight into these relationships (Figure 3.6B). Binarization of the microbial abundances clarifies these relationships and is the starting point for the construction of a dynamic network model of the gut microbiome. Interestingly, principal component analysis demonstrates that the time series data clusters by experimental treatment group, suggesting that our initial assumption of binary

relationships does not lead to significant information loss (Figures 3.2A and 3.2B).

We analyze the topological and dynamic nature of the gut microbiome, focusing on the effect of clindamycin antibiotic and C. difficile infection on gut microbial community structure. We generate a microbial interaction network and dynamic model based on time-series data from a population of mice. We validate a key edge in this interaction network between *Barnesiella* and *C. difficile* through an *in vitro* experiment. Consistent with the literature, our model affirms that solely inoculating a healthy microbiome with *C. difficile* is insufficient to disrupt the healthy intestinal tract microbiome. Additionally, our results demonstrate that clindamycin treatment has a tremendous effect on the microbiome, greatly reducing many microbial genera, and that during the time *C. difficile* is present, a certain subset of bacteria come to dominate the microbiome (Figures 3.2C, S1, and S2).

Our dynamic network model reveals the steady state conditions attainable by this microbial system, how those steady states are reached and maintained, how they relate to the health or disease status of the mice, and how targeted changes in the network can transition the community from a disease state to a healthy state. Furthermore, we examine genome-scale metabolic network reconstructions of the taxa represented in this community, examine broad metabolic differences between the taxa in the community, and probe how metabolism could—and could not—contribute to the mechanistic underpinnings of the observed interactions.

Network Structure

The first feature that stands out in the inferred interaction network is its clustered structure. Clindamycin has a strong influence on the subnetwork containing the two *Lachnospiraceae* nodes and *Barnesiella*. The other subnetwork contains *C. difficile* and other genera that become abundant during *C. difficile* infection (Figure 3.2D). Also worth noticing are the two contradicting edges in the network, between *Coprobacillus* and *Blautia*, and the self-edges for *Blautia* (Figure 3.2D). These arise from rules in the Boolean model that are context-dependent. Such context-dependent rules can manifest as opposite edge types, depending on the state of other nodes in the network. Context-dependent interactions have been demonstrated in many microbial pairings, and nutritional environments can even be designed to induce specific interaction types [58]. It is possible that subtle environmental changes over the course of the experiment altered conditions in a way that flipped the *Coprobacillus-Blautia* interaction. Because the interaction network is derived from time-series data, it is possible to estimate causality, and therefore, derive a directed graph. A directed network with clear, causative interactions can be used to study community dynamics. This is in contrast with association networks, which are often derived from independent samples, and cannot determine direction of causality [48, 59–61]. Such networks are more limited in utility because they cannot be used to predict system behavior over time, or system responses to perturbations [24, 62]. Note that the inferred network structure represents a set of hypotheses as to potential interactions among genera. Determining whether or not the interactions truly occur requires further experimentation, similar to the experimentation completed to validate the edge between *Barnesiella* and *C. difficile*.

Experimental Validation of *Barnesiella* Inhibition of *C. difficile*

We experimentally validated a key edge in the interaction network, and showed that *Barnesiella* can in fact slow *C. difficile* growth. *C. difficile* was grown alone, in co-culture with *B. intestinihominis*, and in *B. intestinihominis* spent media. *C. difficile* grew more slowly in both co-culture and spent-media conditions. Though moderate, the effect was statistically significant (Figure 3.5B). The fact that *C. difficile* growth rate was inhibited under spent-media conditions indicates that *B. intestinihominis*-mediated inhibition does not require *B. intestinihominis* to sense the presence of *C. difficile*. Further, *C. difficile* growth on *B. intestinihominis* spent media demonstrates that the two species have different nutrient requirements. Whether the reduction in growth rate is a result of nutritional limitations (e.g. *C. difficile* resorts to a less preferred carbon source) is unknown, but unlikely given the AUC data.

The AUC—a summation of the OD over the entire time course—is a measure of the total bacterial density achieved over the course of the experiment. It can be thought of as a single metric combining growth rate and biomass production over time. Examining the AUC for all conditions showed that *C. difficile* AUC did not significantly change between fresh media and spent media (Figure 3.5C). Thus, *C. difficile* was able to produce comparable overall biomass despite a reduction in growth rate, further demonstrating that nutrient availability was sufficient in the spent media condition. The AUC for the co-culture was much lower than expected in a simulated null model (Figure 3.5C). Apparently, in co-culture, the total community biomass production capacity is reduced from what would be expected in a scenario without species interaction. Thus, there is a measurable negative interaction between *B. intestinihominis* and *C. difficile* in co-culture that impacts biomass production.

This can be observed over the full time-course of the co-culture, where the overall density is consistently lower than what would be expected in a null model (Figure 3.5D).

Network Dynamics and Perturbation Analysis

Computational perturbation analysis showed that forced overabundance of *Barnesiella* led to a shift from the "disease" state (clindamycin+*C. difficile* treatment group) to a state highly similar to the original healthy state (loss of *C. difficile*). This result is particularly interesting from a therapeutic design standpoint. In this case, the model results indicate that *Barnesiella* may serve as an effective probiotic. Model-driven analysis can be used to identify candidate organisms for probiotic treatments. Recent work by Buffie *et al.* performed a proof-of-concept study in which they used statistical models to identify candidate probiotic organisms, which were then tested on a murine model of *C. difficile* infection [54]. This model-driven approach can be favorably contrasted with the brute-force experimental approach in which successive combinations of microbes are tested until a curative set is found [56]. The model-driven approach requires far fewer experiments, and saves time and resources. While the computational model presented here differs from that used by Buffie *et al.*, the integration of computational models in probiotic design has been shown to be a feasible, effective approach. Improved tools, such as the perturbation analysis presented here, will surely accelerate the probiotic design process and shorten the path to the clinic.

Metabolic Competition Scores Point towards a Non-metabolic Interaction Mechanism

Genome-scale metabolic network reconstructions can be used to estimate the interactions between microbes in a complex community based purely on genome sequence data. Our use of genus-level metabolic network reconstructions (a union of several species-level reconstructions) may not reflect the unique, species-level interactions and heterogeneity within a community. This higher-level model will only capture broad trends and the possible extent of metabolic capacity within a genus. Furthermore, the draft status of these models precludes the effective application of flux balance analysis (FBA) to estimate interactions among genera. This is due to the established lack of precision in draft reconstructions in predictions of growth rates and substrate utilization patterns [63], and the sensitivity of interaction models to metabolic environment and model structure [58, 64]. Future efforts to infer metabolic interactions using FBA and well-curated metabolic networks could provide deeper insights into specific metabolites that are shared (or competed for) between specific microbial pairs.

The application of competition scores demonstrated here (Figure S5A) could potentially be used to quickly establish a rough expectation (notice the spread of competition scores for the species pairs not connected by a path through the network) for community structure—based solely on genomic information—that can then be tested experimentally. Interestingly, the fact that higher competition score is associated with more positive interactions inferred from the Boolean model relates to previous work that demonstrates that higher competition scores were associated with habitat co-occurrence [46]. In this same work, the authors suggest that this effect is due to habitat filtering; that is, microbes with similar metabolic capabilities tend to thrive in similar environments. It has been shown experimentally that microorganisms from the same environment tend to lose net productivity in batch co-culture, indicating similar metabolic requirements [65]. Thus, it appears that metabolically similar organisms tend to co-locate to similar niches, and over evolutionary time, co-localized organisms tend to develop positive relationships with each other.

Understanding this relationship between competition score and interaction type leads to the conclusion that negative interactions are probably not caused by metabolic competition. Of all the genus competition scores with *C. difficile*, *Barnesiella* showed the second lowest (Figure 3.5A). In other words, *Barnesiella* is among the least likely to share a similar metabolic niche with *C. difficile*, which fits with the broad trend mentioned above. The fact that the competition score between *C. difficile* and *Barnesiella* is so low strongly suggests that the negative interaction between them is due, not to competition for scarce resources (although it does not completely exclude the possibility), but rather to some non-metabolic mechanism. The similarity in reaction content between *Barnesiella* and *Enterococcus* indicates similar network structure (Table S5), and yet, *Enterococcus* does not inhibit *C. difficile* in the inferred interaction network (Figure 3.2D). Either the differences that are present between *Barnesiella* (65 unique reactions) and Enterococcus (36 unique reactions) are hints at the mechanism of interaction, or metabolism does not play a significant role in *C. difficile* inhibition in the environment of the gut. For example, enrichment analysis showed that that, with respect to *Enterococcus*, *Barnesiella* is more highly enriched for d-glutamine and d-glutamate metabolism, pantothenate and CoA biosynthesis and LPS biosynthesis. With respect to *Enterococcus*, *Barnesiella* is less enriched in pyrimidine metabolism, and phenylalanine, tyrosine, and trypto-

phan biosynthesis. The possible role of LPS is discussed further on. The possible roles of these other metabolic pathways in *C. difficile* inhibition is unclear.

There is experimental evidence that *Barnesiella* (and other normal flora) may combat pathogen overgrowth through non-metabolic mechanisms. As a first step, it has been shown that VRE can grow in sterile murine cecal contents—indicating the presence of sufficient nutrition to support VRE—but is inhibited in saline-treated cecal contents—indicating that live flora are needed to suppress VRE growth, and that this suppression is not through nutrient sequestration [66]. Further, the presence of *B. intestinihominis* has been demonstrated to prevent and cure VRE infection in mice [55], and is strongly correlated with resistance to *C. difficile* infection in mice [54]. Clearly, *Barnesiella* plays a key role in pathogen inhibition, and pathogen inhibition can be caused by mechanisms other than nutrient competition.

This non-metabolic mechanism may be direct or indirect (Figure 3.6A). We demonstrated *in vitro* that *B. intestinihominis* can inhibit *C. difficile* growth rate (Figure 3.5C and 3.5D). The fact that *C. difficile* grows on *B. intestinihominis* spent media at all indicates that the metabolic requirements of the two species are different, which is consistent with our computational results supporting the hypothesis that *C. difficile* and *Barnesiella* do not compete metabolically (Figure 3.5B). Further, *C. difficile* is moderately inhibited both in co-culture with *B. intestinihominis* and in *B. intestinihominis*-spent media, indicating a direct mechanism of inhibition. In further support of a direct mechanism, it has been shown that *Clostridium scindens* inhibits growth of *C. difficile* through the production of secondary bile acids [54]. Perhaps *Barnesiella* works through an analogous mechanism *in vivo*, enhancing the moderate inhibition observed *in vitro*.

In support of an additional indirect mechanism of bacterial interaction, Buffie and Pamer, in a recent review, hypothesized that the normal flora (of which *Barnesiella* is a member) may prevent pathogen overgrowth by stimulation of a host antimicrobial response (Figure 3.6A) [67]. Specifically, they point out that *Barnesiella* can activate host toll-like receptor TLR signaling, which activates host antimicrobial peptide production. For example, LPS and flagellin have been shown to stimulate the host innate immune response through toll-like receptor (TLR) signaling and production of bactericidal lectins [68, 69]. *Barnesiella* shows enrichment for LPS biosynthesis pathways (Figure 3.4). However, this mechanism did not seem to be responsible for inhibition of VRE by *Barnesiella* [55]. An indirect, host-mediated mechanism is further supported by the fact that members of the normal gut flora can interact differently with

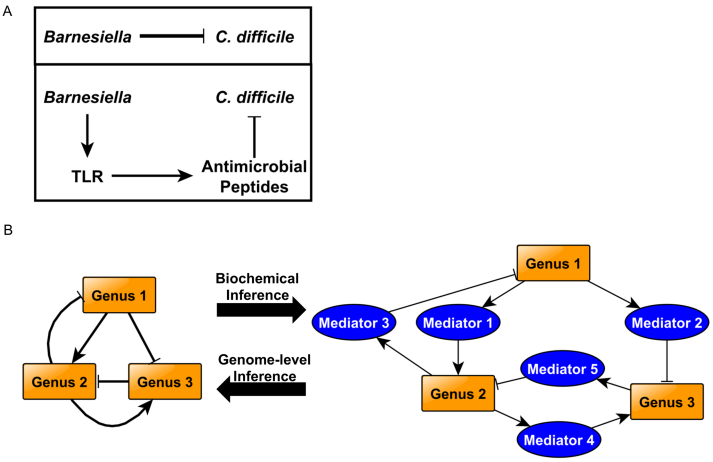


Figure 3.6: **Computational models can bring us closer to true interaction networks.** A) Potential inhibitory mechanisms include direct inhibition of *C. difficile* by *Barnesiella* (e.g. via competition for scarce resources, or toxin production), or indirect inhibition (e.g. through a host antimicrobial response). B) A great deal has been published on the topic of network inference from complex data sets, and more can be done to improve inference methods. Particularly for microbial interaction networks, it is essential to identify, not only the nature of the interactions, but also the underlying mechanisms. Metagenomic genus abundance information can be used to infer causal relationships between bacteria; however, other information sources are required to determine the exact nature of these interactions. Each individual network edge may have very different underlying causes (metabolic, physical interaction, toxin-based, etc.). Including more tools in the pipeline, such as metabolic network reconstructions, bioinformatics tools, etc., will help elucidate these mechanisms, allowing far more rapid hypothesis generation, leading to a more focused effort in the wet lab.

pathogens depending on the host organism [54]. Regardless, any indirect mechanism is in addition to the direct inhibitory mechanism observed *in vitro*. Both direct and indirect mechanisms may play a role *in vivo*, and further work is needed to clearly discern the underlying process that allows *Barnesiella* to play this protective role.

We demonstrate that dynamic Boolean models capture key microbial interactions and dynamics from time-series abundance data in a murine microbiome. We show that this computational approach enables exhaustive *in silico* perturbation, which leads to fast candidate selection for probiotic design. We further describe the use of genome-scale metabolic network reconstructions to explore the metabolic potential attributed to community members, and to estimate metabolic competition and cooperation between members of the microbiome com-

munity. Analysis of genome-scale metabolic network reconstructions indicates that *Barnesiella* likely inhibits *C. difficile* through some non-metabolic mechanism. We present empirical *in vitro* evidence that *B. intestinihominis* does in fact inhibit *C. difficile* growth, likely by a non-metabolic mechanism, and our findings are in good agreement with published results. We present this work as a demonstration of the use of dynamic Boolean models and genome-scale metabolic reconstructions to explore the structure, dynamics, and mechanistic underpinnings of complex microbial communities.

3.7 Acknowledgments

The authors thank Dr. Glynis Kolling (University of Virginia) for help obtaining bacterial isolates and carrying out *in vitro* experiments. The authors further thank Dr. David J. Feith (University of Virginia) for helpful comments/suggestions.

3.8 References

[1] Gordon JI. "Honor thy gut symbionts redux." In: *Science (New York, N.Y.)* 336.6086 (June 2012), pp. 1251–3. DOI: 10.1126/science.1224686.

[2] Bergman EN. "Energy contributions of volatile fatty acids from the gastrointestinal tract in various species." In: *Physiological reviews* 70.2 (Apr. 1990), pp. 567–90.

[3] Rosenberg E, Sharon G, Atad I, and Zilber-Rosenberg I. "The evolution of animals and plants via symbiosis with microorganisms." In: *Environmental microbiology reports* 2.4 (Aug. 2010), pp. 500–6. DOI: 10.1111/j.1758-2229.2010.00177.x.

[4] Kau AL et al. "Human nutrition, the gut microbiome and the immune system." In: *Nature* 474.7351 (June 2011), pp. 327–36. DOI: 10.1038/nature10213.

[5] Faith JJ et al. "The Long-Term Stability of the Human Gut Microbiota". In: *Science* 341.6141 (July 2013), pp. 1237439–1237439. DOI: 10.1126/science.1237439.

[6] Reeves AE et al. "The interplay between microbiome dynamics and pathogen dynamics in a murine model of Clostridium difficile Infection." In: *Gut microbes* 2.3 (), pp. 145–58.

[7] Bartlett JG et al. "Role of Clostridium difficile in antibiotic-associated pseudomembranous colitis." In: *Gastroenterology* 75.5 (Nov. 1978), pp. 778–82.

[8] George WL, Rolfe RD, and Finegold SM. "Clostridium difficile and its cytotoxin in feces of patients with antimicrobial agent-associated diarrhea and miscellaneous conditions." In: *Journal of clinical microbiology* 15.6 (June 1982), pp. 1049–53.

[9] Meyers S et al. "Occurrence of Clostridium difficile toxin during the course of inflammatory bowel disease." In: *Gastroenterology* 80.4 (Apr. 1981), pp. 697–70.

[10] Chang JY et al. "Decreased diversity of the fecal Microbiome in recurrent Clostridium difficile-associated diarrhea." In: *The Journal of infectious diseases* 197.3 (Feb. 2008), pp. 435–8. DOI: 10.1086/525047.

[11] Aroniadis OC and Brandt LJ. "Fecal microbiota transplantation: past, present and future." In: *Current opinion in gastroenterology* 29.1 (Jan. 2013), pp. 79–84. DOI: 10.1097/MOG.0b013e32835a4b3e.

[12] Turnbaugh PJ et al. "An obesity-associated gut microbiome with increased capacity for energy harvest." In: *Nature* 444.7122 (Dec. 2006), pp. 1027–31. DOI: 10.1038/nature05414.

[13] Turnbaugh PJ et al. "A core gut microbiome in obese and lean twins." In: *Nature* 457.7228 (Jan. 2009), pp. 480–4. DOI: 10.1038/nature07540.

[14] Qin J et al. "A metagenome-wide association study of gut microbiota in type 2 diabetes." In: *Nature* 490.7418 (Oct. 2012), pp. 55–60. DOI: 10.1038/nature11450.

[15] Hsiao EY et al. "Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders." In: *Cell* 155.7 (Dec. 2013), pp. 1451–63. DOI: 10.1016/j.cell.2013.11.024.

[16] Dapito DH et al. "Promotion of hepatocellular carcinoma by the intestinal microbiota and TLR4." In: *Cancer cell* 21.4 (Apr. 2012), pp. 504–16. DOI: 10.1016/j.ccr.2012.02.007.

[17] Reddy BS and Watanabe K. "Effect of intestinal microflora on 2,2'-dimethyl-4-aminobiphenyl-induced carcinogenesis in F344 rats." In: *Journal of the National Cancer Institute* 61.5 (Nov. 1978), pp. 1269–71.

[18] Bonneau R et al. "A predictive model for transcriptional control of physiology in a free living cell." In: *Cell* 131.7 (Dec. 2007), pp. 1354–65. DOI: 10.1016/j.cell.2007.10.053.

[19] Schmid AK et al. "The anatomy of microbial cell state transitions in response to oxygen." In: *Genome research* 17.10 (Oct. 2007), pp. 1399–413. DOI: 10.1101/gr.6728007.

[20] Thakar J et al. "Modeling systems-level regulation of host immune responses." In: *PLoS computational biology* 3.6 (June 2007), e109. DOI: 10.1371/journal.pcbi.0030109.

[21] Saez-Rodriguez J et al. "Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction." In: *Molecular systems biology* 5 (2009), p. 331. DOI: 10.1038/msb.2009.87.

[22] Zhang R et al. "Network model of survival signaling in large granular lymphocyte leukemia." In: *Proceedings of the National Academy of Sciences of the United States of America* 105.42 (Oct. 2008), pp. 16308–13. DOI: 10.1073/pnas.0806447105.

[23] Saadatpour A et al. "Dynamical and structural analysis of a T cell survival network identifies novel candidate therapeutic targets for large granular lymphocyte leukemia." In: *PLoS computational biology* 7.11 (Nov. 2011), e1002267. DOI: 10.1371/journal.pcbi.1002267.

[24] Stein RR et al. "Ecological Modeling from Time-Series Inference: Insight into Dynamics and Stability of Intestinal Microbiota". In: *PLoS Computational Biology* 9.12 (Dec. 2013). Ed. by Mering C von, e1003388. DOI: 10.1371/journal.pcbi.1003388.

[25] Marino S et al. "Mathematical modeling of primary succession of murine intestinal microbiota". In: *Proceedings of the National Academy of Sciences* 111.1 (Jan. 2014), pp. 439–444. DOI: 10.1073/pnas.1311322111.

[26] Davidich MI and Bornholdt S. "Boolean network model predicts knockout mutant phenotypes of fission yeast." In: *PloS one* 8.9 (2013), e71786. DOI: 10.1371/journal.pone.0071786.

[27] Bornholdt S. "Systems biology. Less is more in modeling large genetic networks." In: *Science (New York, N.Y.)* 310.5747 (Oct. 2005), pp. 449–51. DOI: 10.1126/science.1119959.

[28] Steinway SN et al. "Network modeling of TGFβ signaling in hepatocellular carcinoma epithelial-to-mesenchymal transition reveals joint sonic hedgehog and Wnt pathway activation." In: *Cancer research* 74.21 (Nov. 2014), pp. 5963–77. DOI: 10.1158/0008-5472.CAN-14-0225.

[29] Naldi A et al. "Cooperative development of logical modelling standards and tools with CoLoMoTo". In: *Bioinformatics* 31.7 (Apr. 2015), pp. 1154–1159. DOI: 10.1093/bioinformatics/btv013.

[30] Campbell C, Yang S, Albert R, and Shea K. "A network model for plant-pollinator community assembly." In: *Proceedings of the National Academy of Sciences of the United States of America* 108.1 (Jan. 2011), pp. 197–202. DOI: 10.1073/pnas.1008204108.

[31] Oberhardt MA, Palsson BØ, and Papin JA. "Applications of genome-scale metabolic reconstructions." In: *Molecular systems biology* 5 (Jan. 2009), p. 320. DOI: 10.1038/msb.2009.77.

[32] Buffie CG et al. "Profound alterations of intestinal microbiota following a single dose of clindamycin results in sustained susceptibility to Clostridium difficile-induced colitis." In: *Infection and immunity* 80.1 (Jan. 2012), pp. 62–73. DOI: 10.1128/IAI.05496-11.

[33] Bartels RH, Beatty JC, and Barskey BA. *Hermite and Cubic Spline Interpolation. An Introduction to Splines for Use in Computer Graphics and Geometric Modelling*. 1998, pp. 9–17.

[34] Saadatpour A and Albert R. "Boolean modeling of biological regulatory networks: a methodology tutorial." In: *Methods (San Diego, Calif.)* 62.1 (July 2013), pp. 3–12. DOI: 10.1016/j.ymeth.2012.10.012.

[35] Berestovsky N and Nakhleh L. "An Evaluation of Methods for Inferring Boolean Networks from Time-Series Data." In: *PloS one* 8.6 (2013), e66031. DOI: 10.1371/journal.pone.0066031.

[36] Lähdesmäki H, Shmulevich I, and Yli-Harja O. "On learning gene regulatory networks under the Boolean network model". In: *Machine Learning* 52.1 (2003), pp. 147–167. DOI: 10.1023/A:1023905711304.

[37] Müssel C, Hopfensitz M, and Kestler HA. "BoolNet–an R package for generation, reconstruction and analysis of Boolean networks." In: *Bioinformatics (Oxford, England)* 26.10 (May 2010), pp. 1378–80. DOI: 10.1093/bioinformatics/btq124.

[38] Chaves M, Albert R, and Sontag ED. "Robustness and fragility of Boolean models for genetic regulatory networks." In: *Journal of theoretical biology* 235.3 (Aug. 2005), pp. 431–49. DOI: 10.1016/j.jtbi.2005.01.023.

[39] Albert I et al. "Boolean network simulations for life scientists." In: *Source code for biology and medicine* 3 (2008), p. 16. DOI: 10.1186/1751-0473-3-16.

[40] Aziz RK et al. "The RAST Server: rapid annotations using subsystems technology." In: *BMC genomics* 9 (2008), p. 75. DOI: 10.1186/1471-2164-9-75.

[41] Overbeek R et al. "The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST)." In: *Nucleic acids research* 42.Database issue (Jan. 2014), pp. D206–14. DOI: 10.1093/nar/gkt1226.

[42] Taffs R et al. "In silico approaches to study mass and energy flows in microbial consortia: a syntrophic case study." In: *BMC systems biology* 3 (Jan. 2009), p. 114. DOI: 10.1186/1752-0509-3-114.

[43] Kanehisa M et al. "Data, information, knowledge and principle: back to metabolism in KEGG". In: *Nucleic Acids Research* 42.D1 (2014), pp. D199–D205. DOI: 10.1093/nar/gkt1076.

[44] Kanehisa M and Goto S. "Kyoto Encyclopedia of Genes and Genomes". In: *Nucleic Acids Research* 28 (2000), pp. 27–30. DOI: 10.1093/nar/28.1.27.

[45] MathWorks. *MATLAB and Statistics Toolbox*. Natick, Massachusetts, USA, 2012.

[46] Levy R and Borenstein E. "Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules." In: *Proceedings of the National Academy of Sciences of the United States of America* 110 (2013), pp. 12804–9. DOI: 10.1073/pnas.1300926110.

[47] Borenstein E, Kupiec M, Feldman MW, and Ruppin E. "Large-scale reconstruction and phylogenetic analysis of metabolic environments." In: *Proceedings of the National Academy of Sciences of the United States of America* 105.38 (2008), pp. 14482–14487. DOI: 10.1073/pnas.0806162105.

[48] Faust K and Raes J. "Microbial interactions: from networks to models". In: *Nature Reviews Microbiology* 10.8 (July 2012), pp. 538–550. DOI: 10.1038/nrmicro2832.

[49] R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria, 2013.

[50] Jensen PA, Dougherty BV, Moutinho TJ, and Papin JA. "Miniaturized Plate Readers for Low-Cost, High-Throughput Phenotypic Screening". In: *Journal of Laboratory Automation* 20.1 (2015), pp. 51–55. DOI: 10.1177/2211068214555414.

[51] Deatherage Kaiser BL et al. "A Multi-Omic View of Host-Pathogen-Commensal Interplay in Salmonella-Mediated Intestinal Infection." In: *PloS one* 8.6 (2013), e67155. DOI: 10.1371/journal.pone.0067155.

[52] Jump RLP et al. "Metabolomics analysis identifies intestinal microbiota-derived biomarkers of colonization resistance in clindamycin-treated mice." In: *PloS one* 9.7 (2014), e101267. DOI: 10.1371/journal.pone.0101267.

[53] Lawley TD et al. "Antibiotic treatment of clostridium difficile carrier mice triggers a supershedder state, spore-mediated transmission, and severe disease in immunocompromised hosts." In: *Infection and immunity* 77.9 (Sept. 2009), pp. 3661–9. DOI: 10.1128/IAI.00558-09.

[54] Buffie CG et al. "Precision microbiome reconstitution restores bile acid mediated resistance to Clostridium difficile". In: *Nature* 517.7533 (2014), pp. 205–208. DOI: 10.1038/nature13828.

[55] Ubeda C et al. "Intestinal microbiota containing Barnesiella species cures vancomycin-resistant Enterococcus faecium colonization." In: *Infection and immunity* 81.3 (Mar. 2013), pp. 965–73. DOI: 10.1128/IAI.01197-12.

[56] Lawley TD et al. "Targeted restoration of the intestinal microbiota with a simple, defined bacteriotherapy resolves relapsing Clostridium difficile disease in mice." In: *PLoS pathogens* 8.10 (2012), e1002995. DOI: 10.1371/journal.ppat.1002995.

[57] Reeves AE, Koenigsknecht MJ, Bergin IL, and Young VB. "Suppression of Clostridium difficile in the gastrointestinal tracts of germfree mice inoculated with a murine isolate from the family Lachnospiraceae." In: *Infection and immunity* 80.11 (Nov. 2012), pp. 3786–94. DOI: 10.1128/IAI.00647-12.

[58] Klitgord N and Segrè D. "Environments that Induce Synthetic Microbial Ecosystems". In: *PLoS Computational Biology* 6.11 (Nov. 2010). Ed. by Papin JA, e1001002. DOI: 10.1371/journal.pcbi.1001002.

[59] Sam Ma Z et al. "Network analysis suggests a potentially 'evil' alliance of opportunistic pathogens inhibited by a cooperative network in human milk bacterial communities." In: *Scientific reports* 5 (2015), p. 8275. DOI: 10.1038/srep08275.

[60] Shankar V et al. "Do gut microbial communities differ in pediatric IBS and health?" In: *Gut microbes* 4.4 (), pp. 347–52. DOI: 10.4161/gmic.24827.

[61] Rigsbee L et al. "Quantitative profiling of gut microbiota of children with diarrhea-predominant irritable bowel syndrome." In: *The American journal of gastroenterology* 107.11 (Nov. 2012), pp. 1740–51. DOI: 10.1038/ajg.2012.287.

[62] Trosvik P, Muinck EJ de, and Stenseth NC. "Biotic interactions and temporal dynamics of the human gastrointestinal microbiota". In: *The ISME Journal* 9.3 (Mar. 2015), pp. 533–541. DOI: 10.1038/ismej.2014.147.

[63] Feist AM et al. "Reconstruction of biochemical networks in microorganisms." In: *Nature reviews. Microbiology* 7.2 (Feb. 2009), pp. 129–43. DOI: 10.1038/nrmicro1949.

[64] Zomorrodi AR and Maranas CD. "OptCom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities." In: *PLoS computational biology* 8.2 (Feb. 2012), e1002363. DOI: 10.1371/journal.pcbi.1002363.

[65] Foster KR and Bell T. "Competition, not cooperation, dominates interactions among culturable microbial species." In: *Current biology : CB* 22.19 (Oct. 2012), pp. 1845–50. DOI: 10.1016/j.cub.2012.08.005.

[66] Pultz NJ et al. "Mechanisms by which anaerobic microbiota inhibit the establishment in mice of intestinal colonization by vancomycin-resistant Enterococcus." In: *The Journal of infectious diseases* 191.6 (Mar. 2005), pp. 949–56. DOI: 10.1086/428090.

[67] Buffie CG and Pamer EG. "Microbiota-mediated colonization resistance against intestinal pathogens." In: *Nature reviews. Immunology* 13.11 (Nov. 2013), pp. 790–801. DOI: 10.1038/nri3535.

[68] Brandl K et al. "Vancomycin-resistant enterococci exploit antibiotic-induced innate immune deficits." In: *Nature* 455.7214 (Oct. 2008), pp. 804–7. DOI: 10.1038/nature07250.

[69] Kinnebrew MA et al. "Bacterial flagellin stimulates Toll-like receptor 5-dependent defense against vancomycin-resistant Enterococcus infection." In: *The Journal of infectious diseases* 201.4 (Feb. 2010), pp. 534–43. DOI: 10.1086/650203.

Chapter 4

Review: Metabolic Network Modeling of Microbial Communities

The text for this chapter has been previously published as a review article here:

Biggs MB, Medlock GL, Kolling GL, Papin JA. (2015). Advanced Review: Metabolic Network Modeling of Microbial Communities. *WIREs Syst Biol Med.* doi: 10.1002/wsbm.1308.

4.1 Context

At the time we wrote this review, there was a lot of interest in using genome-scale metabolic networks to model the metabolic behaviors of microbial communities, but there was no centralized resource which summarized what had already been achieved. Our review filled an important niche in the field, which is highlighted by the fact that it was among the top ten most accessed reviews in WIREs Systems Biology and Medicine in 2015. Completing this review was important for my research in that the reading and synthesis required to write it gave me a more comprehensive perspective on the role of genome-scale metabolic networks in community modeling and how such network models could be integrated with other data types and modeling frameworks.

4.2 Synopsis

Genome-scale metabolic network reconstructions and constraint-based analyses are powerful methods that have the potential to make functional predictions about microbial communities.Genome-scale metabolic networks are used to characterize the metabolic functions of microbial communities via several techniques including species compartmentalization, separating species-level and community-level objectives, dynamic analysis, the 'enzyme-soup' approach, multiscale modeling, and others. There are many challenges in the field, including a need for tools that accurately assign high-level omics signals to individual community members, the need for improved automated network reconstruction methods, and novel algorithms for integrating omics data and engineering communities. As technologies and modeling frameworks improve, we expect that there will be corresponding advances in the fields of ecology, health science, and microbial community engineering.

4.3 Introduction

Microbial communities represent a gargantuan force of nature that exerts influence on global geochemical cycles [1], agriculture [2], human health [3], food preparation [4], and a host of relevant aspects of life on earth [5, 6]. Traditional microbiology has made great strides over the last century in describing and categorizing these microscopic neighbors. More recently, advances in sequencing technologies have provided the first glimpses at the composition of natural microbial communities, including insights into the physiology of non-culturable microbes [7]. Databases are filling with mountains of genomic fragments, gene and protein expression data, and other such large-scale '-omics' information, all describing the content of diverse microbial communities [8, 9]. Despite the plethora of data, we yet lack true understanding of the mechanisms that cause communities to function and interact with their environments [10]. Considering the importance of microbial communities to many global ecosystems, health, and various industries, there is a great need to move beyond a descriptive 'parts list' approach of the field, and transition to more functional, predictive models of microbial community structure and function.

Predictive community models have the potential to engender many beneficial technologies including: rational probiotic design for restoring a diseased intestinal microbiota [11], efficient chemical-producing consortia [12], or optimal bioremediation communities [13]. Furthermore, predictive models will allow novel exploration of basic questions in microbial ecology [14, 15], leading to new insights into the development and evolution of microbial communities (Figure 4.1) [10]. All of these potential applications will require improvements in the mathematical toolbox used to represent biochemical networks and their interactions.

Genome-scale metabolic network reconstructions (GENREs) have been successfully applied to the representation, study, and engineering of single microbes

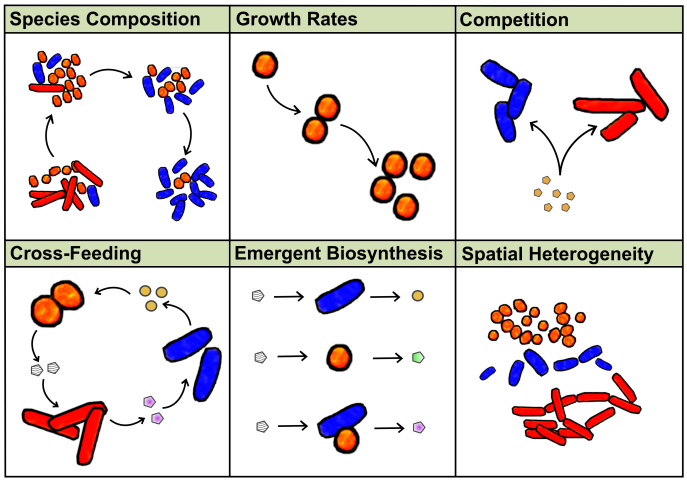


Figure 4.1: There are many aspects of life in a microbial community that would be useful to capture using mathematical models. Techniques utilizing constraint-based metabolic models (sometimes in conjunction with other modeling approaches) are capable of capturing all of these scenarios.

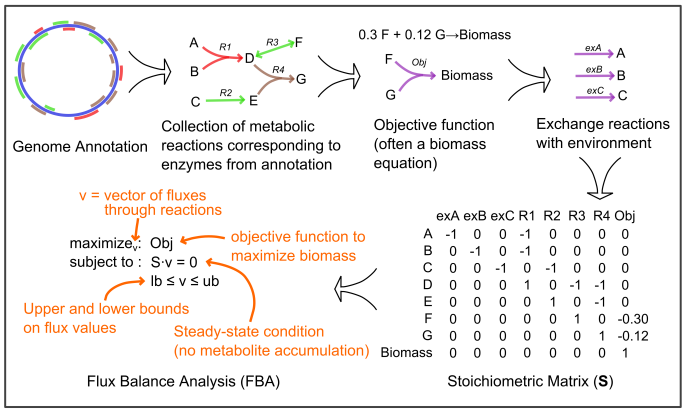


Figure 4.2: **A simple workflow for genome-scale metabolic network reconstruction and accompanying constraint-based analysis**. The process begins with an annotated genome. The metabolic network is derived from this genome annotation by searching databases for homologous proteins with known enzymatic activity. The corresponding metabolic reactions are collected into a draft network reconstruction. This simple procedure can be augmented through gap filling, and often manual curation. A metabolic objective is defined, which for microbes is often assumed to be a biomass equation (i.e., it is assumed that cells are configured to grow as fast as possible). Exchange reactions are defined to allow metabolites to enter and leave the network. All reactions are compiled into a stoichiometric (S) matrix. FBA is a common analytical approach that searches for a flux distribution through the network that optimizes the metabolic objective subject to steady-state constraints and flux bounds.

(Figure 4.2) [16]. The last decade has seen extensive tool development for the analysis of models encompassing single strains up to complex microbial communities [10, 17–22]. Since the first published community model in 2007 of a mutualistic microbial community, the accumulating body of work has highlighted many unique challenges related to microbial community modeling [23]. In this review, we discuss the existing frameworks that have been developed using GENREs for community analysis (Figure 4.3 and Table 4.1), the types of questions that can be addressed, and challenges in the field that present opportunities for progress.

4.4 Current State of the Field

Genome-Scale Metabolic Reconstructions: What They Are and What They Can Do

GENREs are an organized collection of the metabolic reactions that can occur within a biological system. This collection of reactions is inferred from genome annotations, and the resulting gene-to-protein-to-reaction mapping allows genotypephenotype predictions (Figure 4.2). The heart of a GENRE is the stoichiometric (S) matrix, which consists of the stoichiometric coefficients for each reaction represented in the network reconstruction. Mass and charge are balanced for every reaction. This simple representation can be used to explore the space of possible biochemical conversions that can be carried out by the set of reactions in the GENRE. Optimization techniques such as flux balance analysis (FBA) are used to estimate optimal yields given a particular metabolic environment and GENRE [24]. The reconstruction and analysis of GENREs for single organisms have been reviewed extensively [16, 25].

The basic principles for the generation and analysis of GENREs learned from studies of single organisms have been extended in innovative ways to represent the interactions between multiple species within communities. Each of these extensions provides a unique approach to a set of field-specific questions:

1. Structure: How should species models be 'linked' together? Should there be an unbiased, wholesale sharing of metabolites, essentially ignoring species boundaries? Or should metabolite sharing be restricted to only those compounds for which there is empirical evidence? Should species models be linked to the environment/host? And if so, how?

2. Analysis: Should optimization be used to estimate optimal species yield or optimal community yield?

Is there tension between these two objectives *in silico* or *in vitro*? What optimization strategies lead to the most useful analyses?

3. Refinement: How much curation effort should go into the individual species-level reconstructions?

4. Validation: What kind of experimental data can be used to validate model predictions?

5. Applications: Given the model structure and analysis, how informative are the model predictions? How can the model be used to answer impactful questions or inform engineering design choices?

The answers to these questions depend entirely on the purpose of the study, and the questions being asked. We review the various efforts to understand microbial communities using GENREs, and describe each modeling framework utilized to-date.

Compartmentalization

The first framework devised for linking GENREs together is an extension of the compartmentalization approach used for eukaryotic GENREs (Figure 4.3A) [26]. In eukaryotic models, organelles and other compartments are divided, and reactions specific to each compartment are separated by transport reactions [26]. Along the same lines, multiple species-level GENREs are incorporated into a large "meta-stoichiometric matrix" and transport reactions are explicitly added to enable metabolite flux between species compartments, often with an extracellular compartment inserted as a representation of the local environment.

The first community GENRE was developed to represent the mutualistic interaction between *Desulfovibrio vulgaris* and *Methanococcus maripaludis* [23]. The species GENREs consisted only of reactions in central metabolism and were linked using a compartmentalization approach, with shared byproducts and exchange reactions flowing through a shared compartment. FBA was used to estimate optimal growth rate and metabolite fluxes, where the objective function was chosen to be a linear, weighted combination of the biomass functions for each species (with the weights based on experimentally determined species biomass ratios in active communities). Several results highlight the types of questions that can be addressed using this modeling framework. First, FBA results were closest to experimental results during the active-growth phase, where there were no nutrient limitations, which is consistent with the explicit assumption of pseudo-steady-state growth in FBA. Next, simulated flux patterns of primary metabolites matched experimental measurements, such as the large flux in *D. vulgaris* from lactate (the sole carbon substrate) to acetate, with some CO_2 and formate production, and production of a reduced compound, either hydrogen or formate. Flux through *M. maripaludis* showed consumption of acetate or CO_2 and production of CH_4. Further, *in silico* simulations offered insight into the amount of non-productive ATP hydrolysis required to match experimentally measured biomass. This study highlights strengths of the compartmentalization and optimization-based approaches. It allowed the exploration of theoretical limits on growth and nutrient fluxes as a function of the metabolic network structure.

Other studies have successfully utilized the same compartmentalization approach. In one study, this approach was used to computationally design media conditions that induce commensalism or mutualism between microbe pairs [27]. In another, three GENREs (*Bacteroides thetaiotamicron*, *Eubacterium rectale*, and *Methanobrevibacter smithii*) were used to explore the impact of the gut microbiome on host metabolism [28]. To accomplish this, two optimization frameworks were defined which are referred to as the α-problem and the β-problem. The α-problem is used to predict the uptake and secretion of metabolites when the diet and species abundances are known. The β-problem is the inverse, where the model predicts species abundances when metabolite uptake and secretion rates are known. The results from this novel analytical approach are validated using experimental data from gnotobiotic mice. Finally, a host-pathogen interaction between *Mycobacterium tuberculosis* and an alveolar macrophage was simulated using this compartmentalization strategy [29]. The GENRE for *M. tuberculosis* was included as a compartment within the macrophage model, effectively representing a specific metabolic state that *M. tuberculosis* can inhabit during infection [29]. Several other groups have published models that utilize this compartmentalization approach, summarized in Table 4.1.

The compartmentalization framework is an intuitive and simple way to represent microbial interactions. This approach has been used more frequently than any other, providing mechanistic insight into community metabolism and good agreement with experiment. However, the compartmentalization strategy may limit the types of analyses that can be performed. First, this representation of a community inherently forces an assumption of balanced growth making it difficult to account for metabolite accumulation in the environment because of steady-state constraints in FBA. Second, single-level optimization-based analyses of compartmentalized models often assume that each species in the community is growing optimally (i.e., the objective function in FBA is often assumed to be a combination of the objective

A Timeline for Computational Metabolic Systems Biology of Microbial Communities		
C	Stolyar *et al.* (2007)	Recapitulate mutualistic interaction between *Desulfovibrio vulgaris* and *Methanococcus maripaludis* [23]
OM	Christian *et al.* (2007)	Emergent biosynthetic capacity for 99,681 species pairs determined by network expansion [30]
C,ES	Taffs *et al.* (2009)	Interaction of three microbial guilds. Interrogated using three modeling approaches, including simple compartmentalization, "enzyme soup", and an approach based on elementary mode analysis (EMA) [31]
C	Bordbar *et al.* (2010)	*Mycobacterium tuberculosis* embedded in alveolar macrophage metabolic reconstruction [29]
C	Klitgord & Segrè (2010)	Computationally design media to induce commensal and mutualistic interactions between several pairs of species [27]
OM	Sun *et al.* (2010)	Comparative analysis of *Pelobacter carbinolicus* and *Pelobacter propionicus* [32]
C	Freilich *et al.* (2011)	Prediction of competitive and cooperative potential among 6,903 species pairs [33]
DA	Hanly & Henson (2011)	Optimization of glucose/xylose utilization by mixed cultures of *E. coli* and *Saccharomyces cerevisiae* [34]
DA	Zhuang *et al.* (2011)	Simulation of community responses to nutrient modulation. Community included *G. sulfurreducens* and *R. ferrireducens* [35]
DA	Tzamali *et al.* (2011)	Exploration of interactions between many *E. coli* gene-knockout strains [36]
CO	Zomorrodi & Maranas (2012)	OptCom method introduction and analysis of *D. vulgaris* and *M. maripaludis* community [37]
C	Heinken *et al.* (2013)	Interaction of *Bacteroides thetaiotamicron* and mouse host [38]
C	Khandelwal *et al.* (2013)	Development of tools to estimate species abundances and yields, applied to co-culture of *Escherichia coli* auxotrophs [39]
C	Nagarajan *et al.* (2013)	Electron flow between *Geobacter metallireducens* and *Geobacter sulfurreducens*, integrating multi-omics data [40]
C	Shoaie *et al.* (2013)	Interactions between combinations of 2 and 3 of *B. thetaiotamicron*, *Eubacterium rectale*, and *Methanobrevibacter smithii*. Authors also apply OptCom and compare results [28]
DA	Hanly & Henson (2013)	Optimization of glucose/xylose utilization in community of *S. cerevisiae* and *Scherffersomyces stipitis* [41]
OM	Levy & Borenstein (2013)	Analysis of competition and cooperation among all pairs of 154 species using graph-based method [42]
OM	Bartell *et al.* (2014)	Comparative analysis of *Burkholderia cenocepacia* and *Burkholderia multivorans* [43]
OM	Vinay-Lara *et al.* (2014)	Comparative analysis of two strains of *Lactobacillus casei* [44]
CO	El-Semman *et al.* (2014)	Interaction of *Bifidobacterium adolescentis* and *Faecalibacterium prausnitzii*. OptCom and classic FBA are used in analysis [45]
CO,DA	Zomorrodi *et al.* (2014)	OptCom is adapted to dynamic simulations. Simulated communities of *E. coli* auxotrophs, and a uranium-reducing community involving *Geobacter sulfurreducens*, *Rhodoferax ferrireducens*, and *Shewanella oneidensis* [46]
DA	Chiu *et al.* (2014)	Screening of 6,670 two-species communities for emergent biosynthetic capacity [47]
DA	Harcombe *et al.* (2014)	Spatial element integrated with dFBA to model interaction of *Methylobacterium extorquens*, *E. coli*, and *Salmonella enterica* [48]
C	Ye *et al.* (2014)	Analysis of cross-feeding in vitamin-C-producing community composed of *Ketogulonicigenium vulgare* and *Bacillus megaterium* [49]
ES	Tobalina *et al.* (2015)	Analysis of a naphthalene-degrading community [50]

Table 4.1: **A Timeline for Computational Metabolic Systems Biology of Microbial Communities**. C indicates the compartmentalization approach. CO indicates the community objectives method. DA indicates the dynamic analysis approach. ES indicates the enzyme soup approach. OM indicates other methods including graph-based approaches, network expansion, and the comparative method.

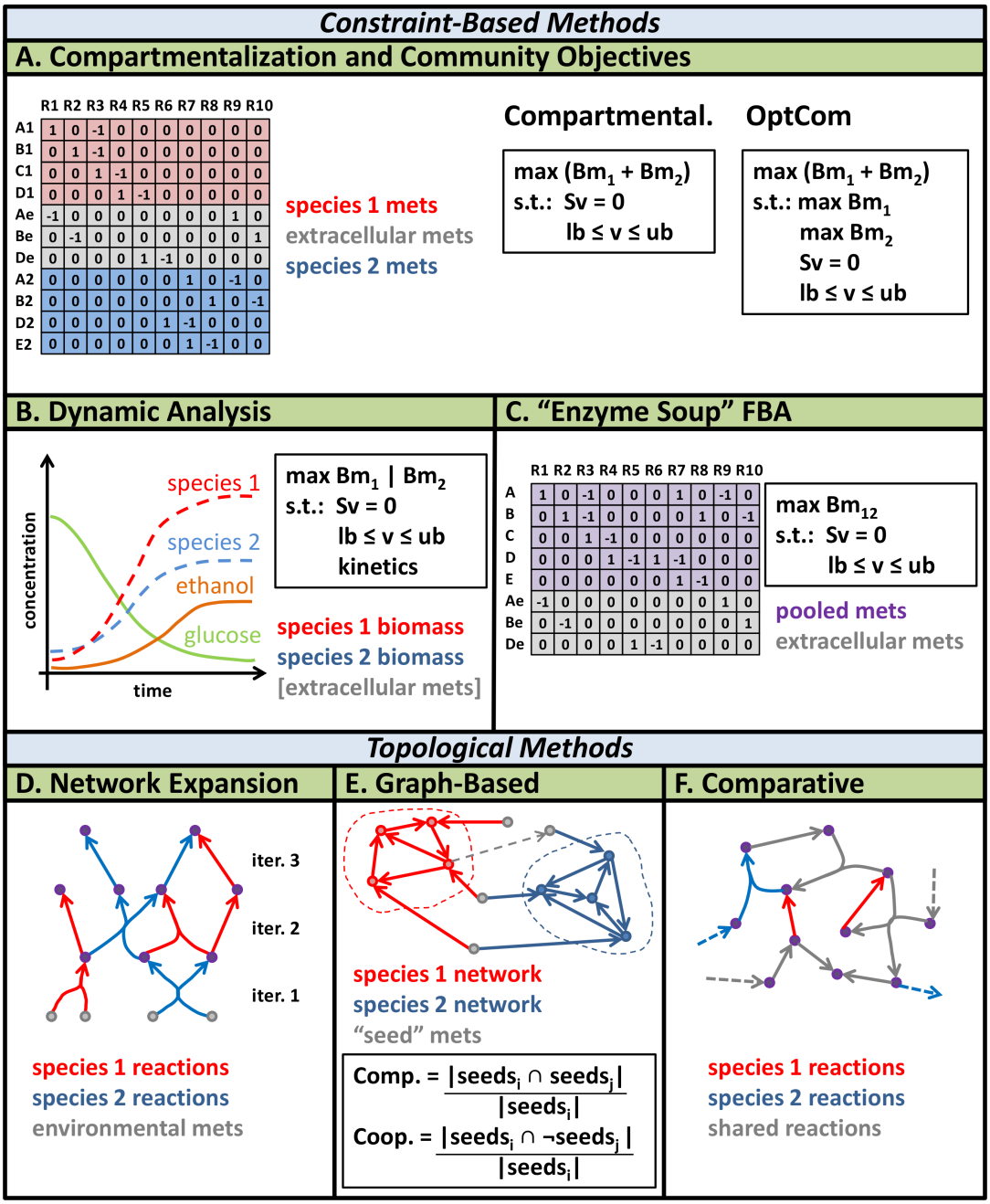


Figure 4.3: **Community modeling frameworks that feature GENREs.** (A) The compartmentalization approach unites all species-level GENREs into a unified stoichiometric matrix with a shared compartment. The objective function is generally assumed to be a linear combination of the individual biomass functions from each species. The community objectives approach (OptCom) is an extension of the simple compartmentalization approach that utilizes a nested, bi-level optimization framework. The bi-level optimization enables the representation of more classes of interactions between species, but comes at the cost of increased computational complexity. (B) Dynamic analysis simulates changes in metabolites and biomass over time, which requires constraints on uptake reaction kinetics. (C) "Enzyme soup" FBA ignores species boundaries and assumes that all reactions can interact in a community-level meta-GENRE. Other methods include: (D) network expansion, which has been used to identify potential emergent biosynthetic capacity between species by comparing species-specific "reachable" metabolites to the result of pooling reactions from both species; (E) graph-based methods, which can be used to quantify general characteristics of an interaction between species, such as the level of expected competition or cooperation; (F) comparative analyses, which are used to assess the differences in gene essentiality, biosynthetic capacity, and resource utilization between species. Note that mets signifies metabolites, Bm stands for biomass, s.t. means subject to, ub and lb signify upper and lower bounds, respectively.

functions for each species). Third, species abundance is assumed to be fixed, rather than allowing for changes in abundance as interactions unfold. These and other limitations are addressed in more recent analytical frameworks.

Community Objectives

The OptCom approach is an extension of the basic compartmentalization strategy that allows for a community-level objective function [37] (as opposed to only considering the species-level objective functions as described above [23]) (Figure 4.3A). A nested, bi-level optimization framework enables the simulation of several classes of metabolic interactions, including mutualism, synergism, commensalism, parasitism, or competition. For example, a mutualistic interaction can be represented by setting the outer optimization problem to maximize the biomass of two interacting community members, subject to the inner optimization conditions for each species. The inner optimization conditions can be customized, and may include maximization of biomass production or alternative objective functions and steady-state constraints. However, a parasitic interaction may be better represented by setting the community objective function to only maximize parasite network biomass production [37]. Beyond the flexibility to represent many qualitatively different types of interactions, OptCom offers a powerful way to think about community interactions: as a result of competing objectives between all community members.

Given the bi-level structure of OptCom, a distinct advantage of this framework is the ability to explore trade-offs between individual and community objectives. A hypothetical example could be that two species maintain suboptimal metabolic states allowing them to catabolize disparate carbon sources and share the resulting byproducts. OptCom is an excellent tool for exploring and explaining such trade-offs between individual metabolic states and community-level optimality. In summary, by altering the community objective and the constraints on interspecies fluxes as part of the outer problem, OptCom can be used to explore many types of communities, and the reasons for observed interactions with respect to trade-offs between objectives. In a representative study, OptCom was used to simulate the interaction between two gut bacteria—*Bifidobacterium adolescentis*, an acetate producer, and *Faecalibacterium prausnitzii*, an acetate consumer and butyrate producer [45]. Flux variability analysis (FVA) was utilized to explore the possible range of flux values for shared metabolites (such as acetate) [45]. OptCom is computationally expensive and not appropriate for some optimization solvers due to the nonlinear constraints [45]. In addition, the analysis is sensitive to the user-defined optimization functions and flux constraints. Therefore, OptCom may be less suitable for poorly defined communities where the metabolic interactions are less well known. Studies that have used OptCom are summarized in Table 4.1.

Dynamic Analysis

Standard FBA results in a set of fluxes—or metabolite consumption/production rates—across a GENRE during pseudo-steady state conditions. In dynamic FBA (dFBA) these fluxes are integrated over time (using standard numerical integration techniques) [51]. With dFBA, it is possible to simulate changes to initial conditions over time, including the consumption and production of metabolites, changes in biomass, and shifts in metabolism in response to environmental changes. dFBA provides an entire time course, as opposed to a single snapshot from standard FBA. Kinetic parameters, particularly relating to uptake rates of limiting metabolites (such as glucose and oxygen) [34, 51] are required for the implementation of dFBA. Challenges with the implementation of dFBA include an increased computational load and a paucity of the required kinetic parameters for many systems (Figure 4.3B).

In the dFBA framework, metabolites are free to accumulate or disappear. Species abundance and metabolic states are free to change in response to interactions and changing environment. Thus, the need for defining a community objective function and to set proper bounds on interspecies fluxes is obviated, given that the proper kinetic parameters are known. Furthermore, it is possible to extend other community modeling methods—including OptCom—and perform dynamic analysis [46].

As an example of multispecies dFBA the co-culture of *Escherichia coli* and *Saccharomyces cerevisiae* was modeled [34]. Each microbe was capable of consuming a unique sugar (glucose or xylose), and the simulations were used to optimize community ethanol production. A similar approach was taken to model the co-culture of *S. cerevisiae* and *Scherffersomyces stipitis* in which the production and degradation of growth-inhibitory compounds such as furfural were represented and growth conditions for ethanol production were optimized [52].

dFBA has been used to identify emergent biosynthetic capacity in 6,670 unique two-species communities [47]. FBA was used to estimate microbial growth at each time point, constrained by kinetic uptake parameters for limiting nutrients [47]. Exchange fluxes can take on a range of possible values, so the lower bound of each was determined by FVA and the sum of all exchange reaction fluxes was minimized. In this way, a reproducible

time course was produced by each simulation. *In silico* species could share metabolites through the shared environment, and emergent metabolites were those that could be produced by a co-culture simulation, but not by either species individually [47]. Interestingly, there is a clear window of phylogenetic distances in which two interacting species are more likely to exhibit emergent metabolic capacity [47].

Another use of dFBA is to capture spatiotemporal dynamics in a community of organisms. In one case, dFBA was used to model the formation of *Pseudomonas aeruginosa* biofilm, where the 'community' is the collection of cells in different metabolic states [53]. In this study, a GENRE corresponding to *P. aeruginosa* was used in a dFBA framework to estimate metabolite secretion, production, and diffusion over time across compartments in an agent-based model of biofilm, recapitulating known features of biofilm formation such as oxygen-limited biofilm growth. In a similar study, dFBA was used to model the spatio-temporal dynamics of three-species microbial communities on a 2D surface [48]. dFBA was used to estimate biomass production, and nutrient concentration changes in local compartments where diffusion allowed changes to impact connected compartments. The authors report successfully predicting the steady-state species composition of an engineered three-member community [48].

These examples demonstrate the versatility of dFBA for modeling small, well-characterized communities, performing large-scale surveys of many potential communities and the possible emergent properties among them, and accounting for spatial dynamics. dFBA represents an exciting and underexplored area of GENRE analysis. dFBA may present a community modeling option with fewer up-front assumptions. However, use of dFBA may be constrained by computational limitations, given the inherent increase in computation over a time-course [47] or spatial environment [48]. Furthermore, dFBA relies on additional kinetic parameters, which may nullify a primary advantage of FBA-based techniques which avoid extensive parameterization. A summary of dFBA-based community models is presented in Table 4.1.

It is worthwhile to compare dFBA with other established dynamic models of microbial communities. The Activated Sludge Model (ASM) has a long history in bioreactor control for wastewater treatment [54]. These models are based on ordinary differential equations, and predict the changes in nutrient concentrations and microbial abundances over time. Nutrients of interest are grouped (e.g., carbon, nitrogen, and phosphorus sources), and microbes grouped according to nutrient utilization (e.g., nitrifying bacteria, phosphorus-accumulating bacteria). An ASM can predict the change in abundance of each microbial group as a function of nutrient concentrations, and subsequent changes in nutrient concentrations as a result of microbial growth. The parameters for these models are chosen to fit experimental data. An ASM does not represent each taxon individually, nor does it account for the metabolic differences between organisms within a group. For example, two bacteria from different taxa may both be classified as "nitrifying". They would likely have different overall metabolic networks, and therefore respond differently to changes in carbon and phosphorus concentrations [55]. An ASM would not account for these taxon-specific metabolic differences, thus resulting predictions may be misleading. In contrast, modeling frameworks that can capitalize on genome-scale metabolic reconstructions, such as dFBA, are capable of accounting for these metabolic differences and can potentially improve prediction accuracy. In addition, dFBA dynamics are typically a function of specific uptake rates while ASM dynamics are a result of fitting the system kinetics to observed data. A recent review describes ASM models as well as other alternative community modeling frameworks that are not based on metabolic network reconstructions [56].

"Enzyme Soup" FBA

In contrast to the simple compartmentalization approach, OptCom, and dFBA, the "enzyme soup" approach [18] ignores species boundaries entirely (Figure 4.3C). The emphasis is on exploration of the metabolic potential of an entire community rather than the interactions between species within a community. A community-level 'enzyme soup' GENRE is produced by annotating a meta-omic dataset for enzyme presence, and the associated reactions are agglomerated into a single set without an attempt to segregate reactions by species. In this framework, any reaction from any species can potentially connect with any other reaction into a "meta-pathway". Early work on this approach ignored stoichiometric constraints within this network, and examined the topological differences between networks reconstructed from healthy and diseased metagenomic data [57].

More recent work pioneered the analysis of these community-level GENREs using constraint-based methods such as FBA to predict biomass production and substrate utilization [50]. In this work, the authors base their reconstruction on metaproteomic data from a naphthalene-enriched soil microbe dataset. They maintain stoichiometric constraints, and assign metabolic activity to taxa within the community based on the taxonomic annotation of the enzymes in the model. The

biomass function is assumed to be a generalized biomass equation borrowed from other organisms, under the assumption that many components of biomass are common to many organisms [50].

The enzyme soup approach has been used successfully to explore the metabolic capacity of complex natural communities. In one study, the interaction of several microbial 'guilds' was studied using both the compartmentalization approach and the enzyme soup method (which they refer to as the "pooled reactions" method), and a "nested consortium analysis" [31]. The guilds represent community functions attributable to prokaryotic oxygenic phototrophs, filamentous anoxygenic phototrophs, and sulfate reducers found in thermophilic, phototrophic mat communities from Yellowstone National Park (USA). The enzyme soup method is appropriate when there is limited *a priori* knowledge about the community; conversely, the compartmentalization approach is a more accurate representation of the biology if extensive knowledge is available about the various community members [31]. The nested consortium analysis is based on elementary mode analysis (EMA) [31, 58]. First, elementary modes (i.e., a pool of valid physiological states) are identified for each guild. Second, elementary modes are re-computed for the entire community using the guild-level modes as input [31]. This "nested consortium analysis" also requires significant *a priori* knowledge of the community in order to select useful elementary modes at the guild-level.

The issue of compartmentalization in GENREs has been discussed in the context of eukaryotic organisms, in which extensive compartmentalization is used to represent organelles and compartments within the cell [26]. For example, analysis of the *S. cerevisiae* GENRE shows that compartmentalization significantly impacts basic properties such as network connectivity and the accuracy of analytical results such as flux values [26]. When considering the "enzyme soup" approach, it is important to consider the loss of accuracy associated with dissolving the boundaries between community members. The main advantage of this approach is the low *a priori* knowledge required, such that it is applicable to little-understood systems. In some sense, the "enzyme soup" strategy can be thought of as placing bounds on the potential metabolic capacity of a microbial community. Further compartmentalization will provide more specific solutions within those bounds. Studies utilizing the enzyme soup method are summarized in Table 4.1.

Other Methods

Other GENRE-based methods of community analysis have been explored, providing yet more vantage points from which to view the metabolism of complex communities. Network Expansion is an algorithm in which the metabolic potential of a set of reactions is explored (Figure 4.3D) [59]. The algorithm starts with a set of metabolites as input (the environment). An initial set of reactions that can use the input metabolites as substrates are added to the network. This network is expanded in subsequent steps as products of the previously added reactions are made available. New reactions that use some part of the accumulating pool of metabolites are added to the network. This approach was extended to a community-level analysis [30]. Given the reaction sets from two organisms, the algorithm assumes that any intermediate metabolites can be shared, and so performs network expansion by pooling the reaction sets from both organisms [30]. This algorithm has been used to identify emergent properties of pairs of microbes, where the combination can produce metabolites that cannot be produced by either parent species [30].

In contrast to the network expansion method [57], approaches based on identifying the "seed set" of a species-level GENRE maintain species-specific boundaries and can be used to estimate metabolic competition or cooperation (Figure 4.3E) [42, 60]. In this graph-based approach, the stoichiometric structure of the GENRE is decomposed to create a directed graph from all substrate-to-product pairs. The resulting graph represents paths between metabolites, but does not contain stoichiometric information. Metabolites that are consumed, but never produced by any reaction in the network, are assigned as network inputs, or the "seed set" [60]. The seed sets for multiple species have been used to estimate the potential for competition or cooperation between species, demonstrating that species tend to co-occur in nature with mutual competitors [42]. These graph-based methods ignore stoichiometry and are therefore not useful for making flux predictions, but rather more generalized statements about network similarity. This approach may be useful when using draft-quality models, for which the accuracy of FBA or similar analyses may be low.

As a final GENRE-based community analysis, we also make note of the comparative approach, which ignores interactions among species and seeks only to identify functional differences between GENREs (Figure 4.3F). For example, a comparative analysis of *Burkholderia cenocepacia* and *Burkholderia multivorans* revealed the unique reactions associated with each, and identified functional outcomes associated with those differences, such as differential virulence factor production capacity [43]. Similarly, a comparative analysis of two strains of *Lactobacillus casei* highlighted functional differences between these industrially-relevant strains [44]. Such com-

parative analyses can help to identify the functional roles of species within large communities by identifying both redundancy and differential metabolic capacity between community members.

Multiscale Models

Outside of the community modeling methods discussed here which are effectively single-scale models (with the possible exception of the spatial models that incorporate dFBA [48, 53]), GENREs have also been successfully incorporated into multiscale models. GENREs of soil microbes have been integrated with a reactive transport model based on local geochemical conditions [61]. A GENRE of the human hepatocyte has been connected with a multicompartment pharmacokinetic model of the human body [62]. Opportunities for multiscale modeling abound with the increasing availability of omics data. For example, in one mouse study the presence of several gut microbes was modulated, and the resulting metabolomics data were fit with a compartmental model that represented the host liver, pancreas, kidney, and adipose tissues [63]. It is clear how a GENRE-based compartmental model could be integrated into an analysis of similar data. These studies highlight the potential gains from combining the strengths of multiple modeling frameworks.

Experimental Data That Meshes with GENRE-Based Community Analysis

GENRE-based community models require data to be constructed, constrained, and validated; for example, quantifying species within a community is important for constraining objective functions (ratios of biomass equations) or validating dFBA simulation results. Many omics technologies are well-suited for this purpose and have been used extensively in GENRE studies of single species. Additional experimental approaches provide insight into microbial community function that may prove useful to GENRE community models in the near future.

Metagenomics answers the question "who is there and what might they be able to do?" The composition of many highly complex microbial communities has been elucidated in recent years through shotgun metagenomic sequencing techniques [7]. This approach can be used to identify the taxa that are present, estimate relative abundances of each taxon, and provide a "parts list" of genes that are present in the community [64]. Metagenomic data can be directly translated into community GENREs using any of the techniques previously discussed; however, difficulties arise during taxonomic assignments within the community. Below, we discuss tools that have been developed to address this issue in metagenomic data. A parallel technology that may eventually overcome challenges in metagenomic sequencing is single-cell sequencing [65, 66]. While throughput is lost, confidence is gained in assignment of function to a distinct organism within the community.

Metatranscriptomics and metaproteomics both help address the question "what are they doing?" Functional data can be used as the basis for constructing a GENRE, or more commonly, to constrain an existing GENRE in order to estimate metabolic pathway usage under specific conditions [40]. Once again, assigning mRNA transcripts or proteins to a specific community member can prove challenging. Single-cell transcriptomics and proteomics can help with this difficulty [67], as can the generation of reference genomes to which transcripts and peptides can be mapped [68].

Other techniques allow the quantification of community member abundances, including real-time PCR, flow cytometry, and in some cases, Coulter counters. Real-time PCR can be used to quantify the abundances of specific community members in a mixed culture [69]. There is generally good correlation between real-time PCR and other measures such as optical density and viability data, unless the culture is under stress, in which case real-time PCR tends to overestimate the number of viable cells [69]. Even so, accurate maximum growth rates can be calculated from real-time PCR regardless of culture conditions [69]. Flow cytometry can be an effective technique to quantify community member abundance, but is limited by the availability of species-specific fluorescent markers [70]. Coulter counters have been used in two-member communities where species vary drastically in size [71].

The burgeoning field of metabolomics offers novel tools to study the anabolic and catabolic capacities of simple and complex communities. Both targeted (measure specific pre-defined metabolites) and non-targeted (measure as many metabolites as possible, without preselection) methods have been applied to studies of microbial communities [72]. Both nuclear magnetic resonance (NMR) and mass spectrometry can be applied in targeted and non-targeted ways, and can provide quantification of metabolite abundances [73–75]. Targeted or non-targeted metabolomic data can be used to constrain, and validate functional outcomes of GENRE simulations. A technique of interest, MALDI-TOF imaging mass spectrometry, is a targeted method capable of measuring metabolite concentrations over a physical space [76, 77]. This technique has been used to identify compounds that are produced during the co-culture of Streptomyces coelicolor with other actinomycetes, and how resulting metabolites localize spatially [76]. This technique may pair particularly well with GENRE mod-

eling techniques that account for spatial information [48, 53].

Many tools for studying microbial communities are applied to the aggregate, which often makes it necessary to partition the data in order to assign activity to a specific community member. Spent media experiments provide a way to dissect cellular interactions within co-culture systems. Using this technique, spent media is produced by growing the first organism in fresh media. The first organism is removed (i.e., by filtration sterilization) resulting in 'sterile' media, and the resulting supernatant is then used to culture a second organism. This technique was used to demonstrate that *Enterobacter cloacae* produced fermentation byproducts that enhanced hydrogen production by *Rhodobacter sphaeroides* [78]. Further, this technique was used to determine the ability of *Lactobacillus* to inhibit growth of *Candida albicans*, *Gardnerella vaginalis*, and *Streptococcus agalactiae* [79]. A disadvantage of this technique is the inability to maintain interspecies signaling, such that the first species (used to produce spent media) cannot respond to the presence of the second species. However, the advantages of this technique are that the direction of interaction and causality can easily be determined, and individual species can be monitored because they are grown separately. In addition, spent-media experiments can be paired with metabolomics and other measures to produce data that can be integrated with GENRE simulations.

Co-culture of community members on solid surfaces such as agar plates can also help to overcome difficulties in partitioning community members. 'Cross streak' analysis is a technique whereby single cultures are mixed on the surface of a plate [80]. Growth and other phenotypes can be visually assessed or paired with other tools such as imaging mass spectrometry [76]. This approach was elegantly used to identify members of the human microbiota that promote growth of antibiotic resistant *Staphylococcus aureus*. At the most basic level, these screens can be used to qualitatively determine the nature of interactions, which can be used to interpret results of GENRE simulations.

4.5 Challenges and Opportunities

Partitioning Communities

Assigning activity to particular species is a fundamental difficulty working with mixed communities and is further compounded when complete genomes are not available for all species in the community [81]. This difficulty assigning activity to a particular species motivates the use of "enzyme soup" methods as discussed above. Considerable progress has been made recently to assemble catalogues of reference genomes, but a great deal remains to be done (and will likely never be "complete") [82]. Traditional genome sequencing has relied on culturing individual isolates in order to extract large quantities of purified DNA, but this is not feasible for the vast majority of organisms [83]. Alternative methods rely on computationally partitioning genomes from mixed metagenomic sequencing samples [81, 83–90]. Three main types of information are used to bin member genomes from mixed samples: (1) DNA composition-based methods, which rely on an empirically observed trend for genomes to display unique "k-mer" frequencies (patterns of one to five bases) [88, 91, 92]; (2) Abundance variations across many samples, where contiguous DNA segments with similar abundance profiles across many samples are likely to originate from the same organism [81, 83–85]; (3) Taxonomic annotations derived from similarity to known taxa [90, 93–95]. There are many active efforts to improve the isolation of individual genetic information from mixed metagenomic samples, and these efforts can translate directly to improved GENRE construction. Species-specific genomes, particularly from non-culturable organisms, will be invaluable resources for understanding the function of complex communities through GENRE-based analysis.

Automating High-Quality Metabolic Reconstructions

Ideally, automated generation of GENREs from metagenomic or genomic data will result in models that have predictive power with minimal manual curation or experimental validation. In practice, however, even the most well-curated GENREs cannot fully recapitulate experimental phenotypes [96]. Therefore, the validity and usefulness of automatically generated GENREs should be assessed by their utility relative to manual reconstructions. A GENRE that can be used to predict growth conditions and gene essentiality will allow a myriad of applications in community modeling and serves as an attainable short-term goal for the development of algorithms for automatically generating GENREs.

Several attempts at automated and semiautomated creation of GENREs have been made, which have been compared and reviewed previously [97, 98]. Many studies report using GENREs created with a combination of automated methods and manual curation [43–45, 99], but there are few, if any, reports of automatically generated GENREs used to contextualize experimental data without manual curation to some degree. Greater precision and throughput is necessary when generating GENREs for uncharacterized, unique communities composed

of diverse sets of microbes that vary greatly across environments or hosts, as is often the case in biomedical or ecological applications.

The next generation of algorithms for automated GENRE generation includes a variety of promising approaches in the context of community modeling. When draft reconstructions are created directly from genome annotations, gap-filling is needed to connect dead-end reactions to produce a functional network. Parsimony-based [100, 101], likelihood-based [102], and phylogeny-based [103] strategies have been developed to fill gaps during automated reconstruction without relying on experimental data. Parsimony-based methods posit that the most parsimonious pathway that fills a gap is the most likely to occur, which results in a smaller GENRE than other gap-filling methods. Likelihood-based methods incorporate multiple gene annotations and use them during the gap-filling stage to present alternative reactions that are each given a likelihood score, greatly expanding the space of possible pathways. While likelihood-based and parsimony-based methods provide similarly accurate results when predicting experimental phenotypes [102], the former provides a framework for finding low-quality gene annotations, which, when removed or fixed, may improve the quality of GENREs created with other methods. Phylogeny-based methods start with the assumption that reactions tend to be more conserved in closely related species than distantly related species. Phylogenetic relationships have been used in the context of gene annotation by assuming that functionally linked proteins have correlated evolution, thus homologs for functionally linked proteins are likely to be present in the same subset of organisms [104]. The outcome of this assumption is that sets of proteins involved in the same function or metabolic pathway can be more accurately annotated in newly sequenced genomes when corresponding homologs involved in that function are identified in another species.

Evolutionary distances have been shown to have a significant, predictable relationship to gene essentiality and growth phenotypes [105]. A framework called CoReCo has been developed which assumes such relationships a priori to enhance the GENRE creation process for multiple species simultaneously [103]. Out of the existing context-based methods, incorporating phylogenetic relationships within a community to guide model creation is particularly interesting because relationships should be obtainable from metagenomic data. Conversely, the assignment of species from metagenomic data could be enhanced by evaluating the function of GENREs created based on multiple putative phylogenies for putative species.

These gap-filling methods present examples of the types of information that need to be integrated in reconstruction algorithms given the constraints of microbial communities. Assumptions based on evolutionary arguments have proven potential in this regard [106], and may have exceptional power that needs to be explored in communities containing both closely and distantly related species. Finally, integrating multiple assumptions and sources of information has the potential to increase GENRE validity in an additive manner and should be explored further.

A final step in model generation that may be particularly relevant to community analyses is model reconciliation. Reconciliation removes the differences between GENREs that represent non-biological noise created through the reconstruction process. Such noise can be due to many factors including, but not limited to, gene annotation uncertainties, differences in the naming conventions in reaction databases used, and unspecified or incorrectly specified reaction reversibility [106]. When reconciliation is performed between models for two related species, the result is typically a reduction in the number of reactions that are unique to each model. This could be particularly useful for searching for therapeutic targets in pathogens that are closely related to a non-pathogen in the same community, as is common in the human microbiome [107]. Reconciliation would result in greater certainty that a target is unique to the pathogen, reducing the probability of off-target effects in commensal organisms.

Reconciliation of automatically generated GENREs from a community may be particularly useful because differences in sequencing quality are likely to be small in a community sample and the same model generation algorithm is likely to be used for all species. The resulting GENREs may be very effective at revealing noise introduced from the reconstruction method, since differences in sequencing and model generation algorithms are controlled. However, reconciliation between two models currently requires a significant amount of manual input and user choice, making it difficult to scale-up to large communities.

Integrating Omics

High-throughput omics technologies such as transcriptomics, proteomics, metabolic flux analysis, and metabolomics all present opportunities for new understanding of microbial communities when integrated with GENRE analysis, but challenges remain with the best approaches for data integration. As with metagenomic information, partitioning omics datasets and assigning them to a particular community member remains a challenge. Along these lines, difficulties may arise when

multiple, highly similar strains exist within the same community interrogated with omics approaches. Transcriptomics and proteomics will both benefit from advances in genome binning and assembly and the resulting species-specific references [82]. Leveraging proteomic techniques, one study used peptide-based ^{13}C metabolic flux analysis to assign metabolic fluxes to species within a community [108]. Metabolomics is perhaps the most challenging, as it can be very difficult to trace the origins of a metabolite in a shared supernatant. Because of inherent limitations in metabolomic technologies that prevent assignment of metabolites to specific community members, GENRE based analyses offer the most effective way to generate *ab initio* hypotheses about the partitioning of metabolic roles within microbial communities.

Regardless of the method of omics partitioning, it is expected that existing methods for omics integration into single-species models will translate well to community models [28, 40]. Many tools for integration of expression data and proteomics data have been developed and validated for individual species [109]; for example, GIMME and MADE represent the trade-off between assumptions and data [110, 111]. GIMME constrains a GENRE with expression data by requiring user-supplied thresholds for each gene, and then optimizing the solution based on consistencies in pathway up/down regulation [110]. MADE is a related algorithm that infers gene-specific thresholds based on multiple expression datasets [111]. If expression data can be easily mapped to reference genomes or proteomes for individual species, these and other existing tools will be applicable. However, in a multispecies community, species abundance is convoluted with gene expression, and precautions should account for such effects. The integration of meta-omics data into community GENRE models may follow a similar path to that of genome assembly algorithms. Metagenomic assembly tools are very similar to single genome assembly tools, with minor changes to address the challenges associated with mixed communities [112, 113]. Perhaps omics data integration algorithms will follow a similar path.

Engineering Communities

Perhaps the most exciting aspect of GENRE community models is the opportunity for community design and engineering. As modeling techniques improve, it is hoped that mechanisms of interspecies interaction will become better understood and more predictable. Early computational tools have already proved valuable from a community engineering standpoint, as demonstrated by the ability to design nutritional environments that modulate the interactions between species in co-culture [27]. FBA-based methods have indicated optimal growth rates of *E. coli* which could be subsequently obtained by adaptive evolution experiments [39, 114]; likewise, it may be possible to use FBA-based methods to predict the necessary individual adaptations of synthetic auxotrophs in co-culture [115]. When large community GENRE models are well-validated, they can be used to explore the impact of specific therapeutic interventions such as prebiotics, probiotics, and targeted removal of species in the human gut microbiome [20, 22, 116]. Analyses such as gene-knockout screens may be extended to species-knockout screens (i.e., sequential removal of each species from the community and analysis of the resulting consequences) [43, 117, 118]. For this type of analysis, it is not currently known what adjustments to flux predictions need to be made to produce accurate simulations. In the case of gene-knockout simulations, techniques such as minimization of metabolic adjustment (MOMA) are used to predict the updated flux distribution [119]. Analogous algorithms will likely be useful for community-level analyses.

It is also unclear how current metabolic engineering tools will be extended to community engineering applications. Algorithms such as OptKnock, Opt-Gene, or the Redirector algorithm have proven useful at a single-organism level [120–122]. Similar optimization frameworks may be devised for community models, but complexity scales not only with the number of species, but also with the many ways species models can be conjoined. There are great incentives to advance such engineering approaches since examples show that microbial consortia are more efficient and robust than a single engineered species [12, 123].

4.6 Conclusion

Metabolic systems biology of microbial communities is an exciting and rapidly developing field with the potential to revolutionize our understanding of microbial communities of societal importance. Existing methods have shown promise, including compartmentalization, separating species-level and community-level objectives, dynamic analysis, the “enzyme-soup” approach, multi-scale modeling, and others. The rise of omics technologies has enabled high-level views of microbial community composition and metabolism, but it remains a challenge to partition community function and assign it to individual community members. Future work is also need to integrate omics data into community-level metabolic models. Moreover, the sheer number of species in many microbial communities demands new automated reconstruction methods that result in GENREs without the

need for further manual curation. As technologies and modeling frameworks improve, we expect that there will be corresponding advances in the fields of ecology, health science, and microbial community engineering.

4.7 References

[1] Rousk J and Bengtson P. "Microbial regulation of global biogeochemical cycles". In: *Frontiers in Microbiology* 5.March (2014), pp. 305–7. DOI: 10.3389/fmicb.2014.00103.

[2] Chaparro JM, Sheflin AM, Manter DK, and Vivanco JM. "Manipulating the soil microbiome to increase soil health and plant fertility". In: *Biology and Fertility of Soils* 48 (2012), pp. 489–499. DOI: 10.1007/s00374-012-0691-4.

[3] Kinross JM, Darzi AW, and Nicholson JK. "Gut microbiome-host interactions in health and disease." In: *Genome medicine* 3 (2011), p. 14. DOI: 10.1186/gm228.

[4] Smid EJ et al. "Functional implications of the microbial community structure of undefined mesophilic starter cultures". In: *Microbial Cell Factories* 13.Suppl 1 (2014), S2. DOI: 10.1186/1475-2859-13-S1-S2.

[5] Van Der Heijden MGa, Bardgett RD, and Van Straalen NM. "The unseen majority: Soil microbes as drivers of plant diversity and productivity in terrestrial ecosystems". In: *Ecology Letters* 11 (2008), pp. 296–310. DOI: 10.1111/j.1461-0248.2007.01139.x.

[6] Holler T et al. "Thermophilic anaerobic oxidation of methane by marine microbial consortia". In: *The ISME Journal* 5 (2011), pp. 1946–1956. DOI: 10.1038/ismej.2011.77.

[7] Handelsman J. "Metagenomics: application of genomics to uncultured microorganisms." In: *Microbiology and molecular biology reviews* 68.4 (2004), pp. 669–685. DOI: 10.1128/MMBR.68.4.669-685.2004.

[8] Meyer F, Paarmann D, D'Souza M, and al. E. "The metagenomics RAST server: a public resource for the automatic phylogenetic and functional analysis of metagenomes". In: *BMC bioinformatics* 9 (2008), p. 386. DOI: 10.1186/1471-2105-9-386.

[9] Hunter S et al. "EBI metagenomics - A new resource for the analysis and archiving of metagenomic data". In: *Nucleic Acids Research* 42.October 2013 (2014), pp. 600–606. DOI: 10.1093/nar/gkt961.

[10] Manor O, Levy R, and Borenstein E. "Mapping the Inner Workings of the Microbiome: Genomic- and Metagenomic-Based Study of Metabolism and Metabolic Interactions in the Human Microbiome." In: *Cell metabolism* 20.5 (Aug. 2014), pp. 742–752. DOI: 10.1016/j.cmet.2014.07.021.

[11] Buffie CG et al. "Precision microbiome reconstitution restores bile acid mediated resistance to Clostridium difficile". In: *Nature* 517.7533 (2014), pp. 205–208. DOI: 10.1038/nature13828.

[12] Brenner K, You L, and Arnold FH. "Engineering microbial consortia: a new frontier in synthetic biology". In: *Trends in Biotechnology* 26.July (2008), pp. 483–489. DOI: 10.1016/j.tibtech.2008.05.004.

[13] Brune KD and Bayer TS. "Engineering microbial consortia to enhance biomining and bioremediation". In: *Frontiers in Microbiology* 3.June (2012), pp. 1–6. DOI: 10.3389/fmicb.2012.00203.

[14] Seth EC and Taga ME. *Nutrient cross-feeding in the microbial world*. 2014. DOI: 10.3389/fmicb.2014.00350.

[15] Morris BEL, Henneberger R, Huber H, and Moissl-Eichinger C. *Microbial syntrophy: Interaction for the common good*. 2013. DOI: 10.1111/1574-6976.12019.

[16] Oberhardt MA, Palsson BØ, and Papin JA. "Applications of genome-scale metabolic reconstructions." In: *Molecular systems biology* 5 (Jan. 2009), p. 320. DOI: 10.1038/msb.2009.77.

[17] Röling WFM, Ferrer M, and Golyshin PN. "Systems approaches to microbial communities and their functioning." In: *Current opinion in biotechnology* 21.4 (Aug. 2010), pp. 532–8. DOI: 10.1016/j.copbio.2010.06.007.

[18] Klitgord N and Segrè D. "Ecosystems biology of microbial metabolism." In: *Current opinion in biotechnology* 22.4 (Aug. 2011), pp. 541–6. DOI: 10.1016/j.copbio.2011.04.018.

[19] Karlsson FH, Nookaew I, Petranovic D, and Nielsen J. "Prospects for systems biology and modeling of the gut microbiome." In: *Trends in biotechnology* 29.6 (June 2011), pp. 251–8. DOI: 10.1016/j.tibtech.2011.01.009.

[20] Borenstein E. "Computational systems biology and in silico modeling of the human microbiome." In: *Briefings in bioinformatics* 13.6 (Nov. 2012), pp. 769–80. DOI: 10.1093/bib/bbs022.

[21] Thiele I, Heinken A, and Fleming RMT. "A systems biology approach to studying the role of microbes in human health." In: *Current opinion in biotechnology* 24.1 (Feb. 2013), pp. 4–12. DOI: 10.1016/j.copbio.2012.10.001.

[22] Shoaie S and Nielsen J. "Elucidating the interactions between the human gut microbiota and its host through metabolic modeling." In: *Frontiers in genetics* 5.April (Jan. 2014), p. 86. DOI: 10.3389/fgene.2014.00086.

[23] Stolyar S et al. "Metabolic modeling of a mutualistic microbial community." In: *Molecular systems biology* 3.92 (2007), p. 92. DOI: 10.1038/msb4100131.

[24] Gianchandani EP, Chavali AK, and Papin JA. "The application of flux balance analysis in systems biology". In: *Wiley interdisciplinary reviews. Systems biology and medicine* 2.3 (2010), pp. 372–382. DOI: 10.1002/wsbm.60.

[25] Thiele I and Palsson BØ. "A protocol for generating a high-quality genome-scale metabolic reconstruction." In: *Nature protocols* 5.1 (2010), pp. 93–121. DOI: 10.1038/nprot.2009.203.

[26] Klitgord N and Segrè D. "The importance of compartmentalization in metabolic flux models: yeast as an ecosystem of organelles." In: *Genome informatics. International Conference on Genome Informatics* 22 (2010), pp. 41–55. DOI: 10.1142/9781848165786_0005.

[27] Klitgord N and Segrè D. "Environments that Induce Synthetic Microbial Ecosystems". In: *PLoS Computational Biology* 6.11 (Nov. 2010). Ed. by Papin JA, e1001002. DOI: 10.1371/journal.pcbi.1001002.

[28] Shoaie S et al. "Understanding the interactions between bacteria in the human gut through metabolic modeling." In: *Scientific reports* 3 (Jan. 2013), p. 2532. DOI: 10.1038/srep02532.

[29] Bordbar A et al. "Insight into human alveolar macrophage and M. tuberculosis interactions via metabolic reconstructions." In: *Molecular systems biology* 6.422 (2010), p. 422. DOI: 10.1038/msb.2010.68.

[30] Christian N, Handorf T, and Ebenhöh O. "Metabolic synergy: increasing biosynthetic capabilities by network cooperation." In: *Genome informatics. International Conference on Genome Informatics* 18 (2007), pp. 320–329.

[31] Taffs R et al. "In silico approaches to study mass and energy flows in microbial consortia: a syntrophic case study." In: *BMC systems biology* 3 (Jan. 2009), p. 114. DOI: 10.1186/1752-0509-3-114.

[32] Sun J et al. "Constraint-based modeling analysis of the metabolism of two Pelobacter species." In: *BMC systems biology* 4.1 (Jan. 2010), p. 174. DOI: 10.1186/1752-0509-4-174.

[33] Freilich S et al. "Competitive and cooperative metabolic interactions in bacterial communities." In: *Nature communications* 2 (Jan. 2011), p. 589. DOI: 10.1038/ncomms1597.

[34] Hanly TJ and Henson Ma. "Dynamic flux balance modeling of microbial co-cultures for efficient batch fermentation of glucose and xylose mixtures". In: *Biotechnology and Bioengineering* 108.2 (2011), pp. 376–385. DOI: 10.1002/bit.22954.

[35] Zhuang K et al. "Genome-scale dynamic modeling of the competition between Rhodoferax and Geobacter in anoxic subsurface environments." In: *The ISME journal* 5.2 (Feb. 2011), pp. 305–16. DOI: 10.1038/ismej.2010.117.

[36] Tzamali E, Poirazi P, Tollis IG, and Reczko M. "A computational exploration of bacterial metabolic diversity identifying metabolic interactions and growth-efficient strain communities." In: *BMC systems biology* 5.1 (Jan. 2011), p. 167. DOI: 10.1186/1752-0509-5-167.

[37] Zomorrodi AR and Maranas CD. "OptCom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities." In: *PLoS computational biology* 8.2 (Feb. 2012), e1002363. DOI: 10.1371/journal.pcbi.1002363.

[38] Heinken A, Sahoo S, Fleming RMT, and Thiele I. "Systems-level characterization of a host-microbe metabolic symbiosis in the mammalian gut." In: *Gut microbes* 4.February 2015 (2013), pp. 28–40. DOI: 10.4161/gutm.22370.

[39] Khandelwal Ra et al. "Community flux balance analysis for microbial consortia at balanced growth." In: *PloS one* 8.5 (Jan. 2013), e64567. DOI: 10.1371/journal.pone.0064567.

[40] Nagarajan H et al. "Characterization and modelling of interspecies electron transfer mechanisms and microbial community dynamics of a syntrophic association." In: *Nature communications* 4 (Jan. 2013), p. 2809. DOI: 10.1038/ncomms3809.

[41] Hanly TJ and Henson Ma. "Dynamic metabolic modeling of a microaerobic yeast co-culture: predicting and optimizing ethanol production from glucose/xylose mixtures." In: *Biotechnology for biofuels* 6.1 (Jan. 2013), p. 44. DOI: 10.1186/1754-6834-6-44.

[42] Levy R and Borenstein E. "Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules." In: *Proceedings of the National Academy of Sciences of the United States of America* 110 (2013), pp. 12804–9. DOI: 10.1073/pnas.1300926110.

[43] Bartell JA et al. "Comparative Metabolic Systems Analysis of Pathogenic Burkholderia". In: *Journal of Bacteriology* 196.2 (2014), pp. 210–226. DOI: 10.1128/JB.00997-13.

[44] Vinay-Lara E et al. "Genome -Scale Reconstruction of Metabolic Networks of Lactobacillus casei ATCC 334 and 12A." In: *PloS one* 9.11 (Jan. 2014), e110785. DOI: 10.1371/journal.pone.0110785.

[45] El-Semman IE et al. "Genome-scale metabolic reconstructions of Bifidobacterium adolescentis L2-32 and Faecalibacterium prausnitzii A2-165 and their interaction." In: *BMC systems biology* 8.1 (Jan. 2014), p. 41. DOI: 10.1186/1752-0509-8-41.

[46] Zomorrodi AR, Islam MM, and Maranas CD. "D-OptCom: Dynamic Multi-level and Multi-objective Metabolic Modeling of Microbial Communities". In: *ACS Synthetic Biology* 3 (2014), pp. 247–257. DOI: 10.1021/sb4001307.

[47] Chiu HC, Levy R, and Borenstein E. "Emergent biosynthetic capacity in simple microbial communities." In: *PLoS computational biology* 10.7 (July 2014), e1003695. DOI: 10.1371/journal.pcbi.1003695.

[48] Harcombe WR et al. "Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics." In: *Cell reports* 7.4 (May 2014), pp. 1104–15. DOI: 10.1016/j.celrep.2014.03.070.

[49] Ye C, Zou W, Xu N, and Liu L. "Metabolic model reconstruction and analysis of an artificial microbial ecosystem for vitamin C production." In: *Journal of biotechnology* 182-183 (July 2014), pp. 61–7. DOI: 10.1016/j.jbiotec.2014.04.027.

[50] Tobalina L et al. "Context-specific metabolic network reconstruction of a naphthalene degrading bacterial community guided by metaproteomic data". In: *Bioinformatics* (2015), pp. 1–9. DOI: 10.1093/bioinformatics/btv036.

[51] Mahadevan R, Edwards JS, and Doyle FJ. "Dynamic flux balance analysis of diauxic growth in Escherichia coli." In: *Biophysical journal* 83.3 (2002), pp. 1331–1340. DOI: 10.1016/S0006-3495(02)73903-9.

[52] Hanly TJ and Henson Ma. "Dynamic model-based analysis of furfural and HMF detoxification by pure and mixed batch cultures of S. cerevisiae and S. stipitis". In: *Biotechnology and Bioengineering* 111.2 (2014), pp. 272–284. DOI: 10.1002/bit.25101.

[53] Biggs MB and Papin Ja. "Novel Multiscale Modeling Tool Applied to Pseudomonas aeruginosa Biofilm Formation". In: *PLoS ONE* 8.10 (2013), pp. 1–8. DOI: 10.1371/journal.pone.0078011.

[54] Hauduc H et al. *Critical review of activated sludge modeling: State of process knowledge, modeling concepts, and limitations*. 2013. DOI: 10.1002/bit.24624.

[55] Oehmen A et al. "Incorporating microbial ecology into the metabolic modelling of polyphosphate accumulating organisms and glycogen accumulating organisms". In: *Water Research* 44 (2010), pp. 4992–5004. DOI: 10.1016/j.watres.2010.06.071.

[56] Bucci V and Xavier JB. *Towards Predictive Models of the Human Gut Microbiome*. 2014. DOI: 10.1016/j.jmb.2014.03.017.

[57] Greenblum S, Turnbaugh PJ, and Borenstein E. "Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease". In: *Proceedings of the National Academy of Sciences* 109 (2012), pp. 594–599. DOI: 10.1073/pnas.1116053109/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1116053109.

[58] Trinh CT, Wlaschin A, and Srienc F. *Elementary mode analysis: A useful metabolic pathway analysis tool for characterizing cellular metabolism*. 2009. DOI: 10.1007/s00253-008-1770-1.

[59] Handorf T, Ebenhöh O, and Heinrich R. "Expanding metabolic networks: Scopes of compounds, robustness, and evolution". In: *Journal of Molecular Evolution* 61 (2005), pp. 498–512. DOI: 10.1007/s00239-005-0027-1.

[60] Borenstein E, Kupiec M, Feldman MW, and Ruppin E. "Large-scale reconstruction and phylogenetic analysis of metabolic environments." In: *Proceedings of the National Academy of Sciences of the United States of America* 105.38 (2008), pp. 14482–14487. DOI: 10.1073/pnas.0806162105.

[61] Fang Y et al. "Direct coupling of a genome-scale microbial in silico model and a groundwater reactive transport model". In: *Journal of Contaminant Hydrology* 122.1-4 (2011), pp. 96–103. DOI: 10.1016/j.jconhyd.2010.11.007.

[62] Krauss M et al. "Integrating cellular metabolism into a multiscale whole-body model." In: *PLoS computational biology* 8.10 (Jan. 2012), e1002750. DOI: 10.1371/journal.pcbi.1002750.

[63] Martin FPJ et al. "Panorganismal gut microbiome-host metabolic crosstalk". In: *Journal of Proteome Research* 8 (2009), pp. 2090–2105. DOI: 10.1021/pr801068x.

[64] Karlsson FH, Nookaew I, and Nielsen J. "Metagenomic Data Utilization and Analysis (MEDUSA) and Construction of a Global Gut Microbial Gene Catalogue". In: *PLoS Computational Biology* 10.7 (2014). DOI: 10.1371/journal.pcbi.1003706.

[65] Lasken RS and McLean JS. "Recent advances in genomic DNA sequencing of microbial species from single cells". In: *Nature Reviews. Genetics* 15.9 (2014), pp. 577–584. DOI: 10.1038/nrg3785.

[66] Woyke T et al. "Assembling the marine metagenome, one cell at a time." In: *PloS one* 4.4 (Jan. 2009), e5299. DOI: 10.1371/journal.pone.0005299.

[67] Taniguchi Y et al. "Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells." In: *Science (New York, N.Y.)* 329.533 (2010), pp. 533–538. DOI: 10.2142/biophys.51.136.

[68] Embree M et al. "Single-cell genome and metatranscriptome sequencing reveal metabolic interactions of an alkane-degrading methanogenic community." In: *The ISME journal* 8 (2014), pp. 757–67. DOI: 10.1038/ismej.2013.187.

[69] Reichert-Schwillinsky F et al. "Stress- and growth rate-related differences between plate count and real-time PCR data during growth of Listeria monocytogenes". In: *Applied and Environmental Microbiology* 75.7 (2009), pp. 2132–2138. DOI: 10.1128/AEM.01796-08.

[70] Besmer MD et al. "The feasibility of automated online flow cytometry for In-situ monitoring of microbial dynamics in aquatic ecosystems". In: *Frontiers in Microbiology* 5.June (2014), pp. 1–12. DOI: 10.3389/fmicb.2014.00265.

[71] Hanly TJ, Urello M, and Henson Ma. "Dynamic flux balance modeling of S. cerevisiae and E. coli cocultures for efficient consumption of glucose/xylose mixtures." In: *Applied microbiology and biotechnology* 93.6 (Mar. 2012), pp. 2529–41. DOI: 10.1007/s00253-011-3628-1.

[72] Sévin DC, Kuehne A, Zamboni N, and Sauer U. "Biological insights through nontargeted metabolomics". In: *Current Opinion in Biotechnology* 34 (2015), pp. 1–8. DOI: 10.1016/j.copbio.2014.10.001.

[73] Walker A et al. "Distinct signatures of host-microbial meta-metabolome and gut microbiome in two C57BL/6 strains under high-fat diet." In: *The ISME journal* (2014), pp. 1–17. DOI: 10.1038/ismej.2014.79.

[74] Kok MGM et al. "Hydrophilic interaction chromatography-mass spectrometry for anionic metabolic profiling of urine from antibiotic-treated rats". In: *Journal of Pharmaceutical and Biomedical Analysis* 92 (2014), pp. 98–104. DOI: 10.1016/j.jpba.2014.01.008.

[75] Gan XT et al. "Probiotic administration attenuates myocardial hypertrophy and heart failure after myocardial infarction in the rat". In: *Circulation: Heart Failure* 7 (2014), pp. 491–499. DOI: 10.1161/CIRCHEARTFAILURE.113.000978.

[76] Traxler MF et al. "Interspecies interactions stimulate diversification of the Streptomyces coelicolor secreted metabolome". In: *mBio* 4.4 (2013). DOI: 10.1128/mBio.00459-13.

[77] Barger SR et al. "Imaging secondary metabolism of Streptomyces sp. Mg1 during cellular lysis and colony degradation of competing Bacillus subtilis". In: *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology* 102 (2012), pp. 435–445. DOI: 10.1007/s10482-012-9769-0.

[78] Nath K, Kumar A, and Das D. "Hydrogen production by Rhodobacter sphaeroides strain O.U.001 using spent media of Enterobacter cloacae strain DM11". In: *Applied Microbiology and Biotechnology* 68 (2005), pp. 533–541. DOI: 10.1007/s00253-005-1887-4.

[79] Seta F, Hunter M, and Larsen B. "In vitro Evaluation of Small Molecule Inhibitors and Probiotic Byproducts on Growth and Viability of Vaginal Microorganisms". In: *British Journal of Medicine and Medical Research* 4.August (2014), pp. 5779–5792. DOI: 10.9734/BJMMR/2014/12327.

[80] Michelsen CF et al. "Staphylococcus aureus Alters Growth Activity, Autolysis, and Antibiotic Tolerance in a Human Host-Adapted Pseudomonas aeruginosa Lineage". In: *Journal of Bacteriology* 196.22 (Nov. 2014), pp. 3903–3911. DOI: 10.1128/JB.02006-14.

[81] Nielsen HB et al. "Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes." In: *Nature biotechnology* 32.8 (2014), pp. 822–828. DOI: 10.1038/nbt.2939.

[82] Human Microbiome Jumpstart Reference Strains Consortium. "A catalog of reference genomes from the human microbiome." In: *Science (New York, N.Y.)* 328.May (2010), pp. 994–999. DOI: 10.1126/science.1183605.

[83] Albertsen M et al. "Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes." In: *Nature biotechnology* 31.6 (2013), pp. 533–8. DOI: 10.1038/nbt.2579.

[84] Alneberg J et al. "Binning metagenomic contigs by coverage and composition". In: *Nature Methods* (2014). DOI: 10.1038/nmeth.3103.

[85] Sharon I et al. "Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization". In: *Genome Research* 23 (2013), pp. 111–120. DOI: 10.1101/gr.142315.112.

[86] Carr R, Shen-Orr SS, and Borenstein E. "Reconstructing the Genomic Content of Microbiome Taxa through Shotgun Metagenomic Deconvolution". In: *PLoS Computational Biology* 9 (2013). DOI: 10.1371/journal.pcbi.1003292.

[87] Alneberg J et al. "CONCOCT: Clustering cONtigs on COverage and ComposiTion". In: *Arxiv preprint arXiv:1312.4038v1* (2013), p. 28. arXiv: 1312.4038.

[88] Gori F, Mavroedis D, Jetten MSM, and Marchiori E. "Genomic signatures for metagenomic data analysis: Exploiting the reverse complementarity of tetranucleotides". In: *2011 IEEE International Conference on Systems Biology, ISB 2011* (2011), pp. 149–154. DOI: 10.1109/ISB.2011.6033147.

[89] Wu YW et al. "MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm". In: *Microbiome* 2 (2014), p. 26. DOI: 10.1186/2049-2618-2-26.

[90] Brady A and Salzberg SL. "Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models." In: *Nature methods* 6.9 (2009), pp. 673–676. DOI: 10.1038/nmeth.1358.

[91] Teeling H et al. "Application of tetranucleotide frequencies for the assignment of genomic fragments." In: *Environmental microbiology* 6.9 (Sept. 2004), pp. 938–47. DOI: 10.1111/j.1462-2920.2004.00624.x.

[92] Kelley DR and Salzberg SL. "Clustering metagenomic sequences with interpolated Markov models." In: *BMC bioinformatics* 11.1 (2010), p. 544. DOI: 10.1186/1471-2105-11-544.

[93] Patil KR et al. "Taxonomic metagenome sequence assignment with structured output models." In: *Nature methods* 8.3 (2011), pp. 191–192. DOI: 10.1038/nmeth0311-191.

[94] MacDonald NJ, Parks DH, and Beiko RG. "Rapid identification of high-confidence taxonomic assignments for metagenomic data". In: *Nucleic Acids Research* 40 (2012). DOI: 10.1093/nar/gks335.

[95] Jiang H et al. "A Statistical Framework for Accurate Taxonomic Assignment of Metagenomic Sequencing Reads". In: *PLoS ONE* 7.10 (2012). DOI: 10.1371/journal.pone.0046450.

[96] Orth JD et al. *A comprehensive genome-scale reconstruction of Escherichia coli metabolism.* 2011. DOI: 10.1038/msb.2011.65.

[97] Hamilton JJ and Reed JL. *Software platforms to facilitate reconstructing genome-scale metabolic networks.* 2014. DOI: 10.1111/1462-2920.12312.

[98] Saha R, Chowdhury A, and Maranas CD. "Recent advances in the reconstruction of metabolic models and integration of omics data." In: *Current opinion in biotechnology* 29C (Mar. 2014), pp. 39–45. DOI: 10.1016/j.copbio.2014.02.011.

[99] Alam MT, Medema MH, Takano E, and Breitling R. "Comparative genome-scale metabolic modeling of actinomycetes: The topology of essential core metabolism". In: *FEBS Letters* 585 (2011), pp. 2389–2394. DOI: 10.1016/j.febslet.2011.06.014.

[100] Satish Kumar V, Dasika MS, and Maranas CD. "Optimization based automated curation of metabolic reconstructions." In: *BMC bioinformatics* 8 (2007), p. 212. DOI: 10.1186/1471-2105-8-212.

[101] Kumar VS and Maranas CD. "GrowMatch: An automated method for reconciling in silico/in vivo growth predictions". In: *PLoS Computational Biology* 5 (2009). DOI: 10.1371/journal.pcbi.1000308.

[102] Benedict MN et al. "Likelihood-Based Gene Annotations for Gap Filling and Quality Assessment in Genome-Scale Metabolic Models". In: *PLoS Computational Biology* 10 (2014), e1003882. DOI: 10.1371/journal.pcbi.1003882.

[103] Pitkänen E et al. "Comparative genome-scale reconstruction of gapless metabolic networks for present and ancestral species." In: *PLoS computational biology* 10.2 (2014), e1003465. DOI: 10.1371/journal.pcbi.1003465.

[104] Pellegrini M et al. "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles." In: *Proceedings of the National Academy of Sciences of the United States of America* 96 (1999), pp. 4285–4288. DOI: 10.1073/pnas.96.8.4285.

[105] Plata G, Henry CS, and Vitkup D. "Long-term phenotypic evolution of bacteria". In: *Nature* (2014). DOI: 10.1038/nature13827.

[106] Oberhardt Ma, Puchałka J, Santos VaPM dos, and Papin Ja. "Reconciliation of genome-scale metabolic reconstructions for comparative systems analysis". In: *PLoS Computational Biology* 7.3 (2011). DOI: 10.1371/journal.pcbi.1001116.

[107] Stecher B et al. "Like will to like: Abundances of closely related species can predict susceptibility to intestinal colonization by pathogenic and commensal bacteria". In: *PLoS Pathogens* 6 (2010). DOI: 10.1371/journal.ppat.1000711.

[108] Ghosh A et al. "A Peptide-Based Method for 13C Metabolic Flux Analysis in Microbial Communities". In: *PLoS Computational Biology* 10.9 (Sept. 2014). Ed. by Ouzounis CA, e1003827. DOI: 10.1371/journal.pcbi.1003827.

[109] Blazier AS and Papin JA. "Integration of expression data in genome-scale metabolic network reconstructions." In: *Frontiers in physiology* 3 (Jan. 2012), p. 299. DOI: 10.3389/fphys.2012.00299.

[110] Becker Sa and Palsson BO. "Context-specific metabolic networks are consistent with experiments". In: *PLoS Computational Biology* 4.5 (2008). DOI: 10.1371/journal.pcbi.1000082.

[111] Jensen Pa and Papin Ja. "Functional integration of a metabolic network model and expression data without arbitrary thresholding". In: *Bioinformatics* 27.4 (2011), pp. 541–547. DOI: 10.1093/bioinformatics/btq702.

[112] Namiki T, Hachiya T, Tanaka H, and Sakakibara Y. "MetaVelvet: An extension of Velvet assembler to de novo metagenome assembly from short sequence reads". In: *Nucleic Acids Research* 40 (2012), pp. 1–12. DOI: 10.1093/nar/gks678.

[113] Afiahayati, Sato K, and Sakakibara Y. "MetaVelvet-SL : An extension of Velvet assembler to de novo metagenomic assembler utilizing supervised learning". In: *DNA Research* (2014), pp. 1–9. DOI: dsu041[pii] 10.1093/dnares/dsu041.

[114] Ibarra RU, Edwards JS, and Palsson BO. "Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth." In: *Nature* 420.November (2002), pp. 186–189. DOI: 10.1038/nature01149.

[115] Wintermute EH and Silver Pa. "Emergent cooperation in microbial metabolism." In: *Molecular systems biology* 6.407 (Sept. 2010), p. 407. DOI: 10.1038/msb.2010.66.

[116] Waldor MK et al. "Where next for microbiome research?" In: *PLoS Biol* (2015), pp. 1–9. DOI: 10.1371/journal.pbio.1002050.

[117] Becker SA et al. "Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox." In: *Nature protocols* 2.3 (Jan. 2007), pp. 727–38. DOI: 10.1038/nprot.2007.99.

[118] Oberhardt MA et al. "Genome-scale metabolic network analysis of the opportunistic pathogen Pseudomonas aeruginosa PAO1." In: *Journal of bacteriology* 190.8 (Apr. 2008), pp. 2790–803. DOI: 10.1128/JB.01583-07.

[119] Segrè D, Vitkup D, and Church GM. "Analysis of optimality in natural and perturbed metabolic networks." In: *Proceedings of the National Academy of Sciences of the United States of America* 99.Track II (2002), pp. 15112–15117. DOI: 10.1073/pnas.232349399.

[120] Burgard AP, Pharkya P, and Maranas CD. "Optknock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization". In: *Biotechnology and Bioengineering* 84.6 (2003), pp. 647–657. DOI: 10.1002/bit.10803.

[121] Patil KR, Rocha I, Förster J, and Nielsen J. "Evolutionary programming as a platform for in silico metabolic engineering." In: *BMC bioinformatics* 6 (2005), p. 308. DOI: 10.1186/1471-2105-6-308.

[122] Rockwell G, Guido NJ, and Church GM. "Redirector: designing cell factories by reconstructing the metabolic objective." In: *PLoS computational biology* 9.1 (Jan. 2013), e1002882. DOI: 10.1371/journal.pcbi.1002882.

[123] Shong J, Jimenez Diaz MR, and Collins CH. "Towards synthetic microbial consortia for bioprocessing." In: *Current opinion in biotechnology* 23.5 (Oct. 2012), pp. 798–802. DOI: 10.1016/j.copbio.2012.02.001.

Chapter 5

Metabolic Network-guided Binning of Metagenomic Sequence Fragments

The text for this chapter has been previously published as a research article here:

Biggs MB and Papin JA. (2016). Metabolic Network-guided Binning of Metagenomic Sequence Fragments. *Bioinformatics*. 32(6):867–874. doi: 10.1093/bioinformatics/btv671.

5.1 Context

This paper started with an idea that sounded interesting but far-fetched. We decided to do some exploratory simulations on toy data, just to see if the idea held water, and we were surprised to find out that it actually worked. Once we knew the idea was viable, I took some time to brainstorm the experiments that would need to be done and outline a possible paper (I credit Phil Yen for being a good example of deliberate planning). From there, the process of doing the work and writing the publication was relatively fast (just a couple of months from start to finish). The experience of writing this paper taught me a couple of key principles: 1) do the make-or-break experiment as early as possible (in this case, we did simulations first thing) and 2) good planning in the beginning makes the whole process more coherent and prevents losing sight of the big picture.

This particular application of genome-scale metabolic network reconstructions is "nifty", but may not find wide application because of recent technologies that reduce the need to computationally bin metagenomic sequence fragments. Single-cell sequencing and long-read sequencing technologies are two that come to mind. Even though SONEC may not be found on every future bioinformatician's hard drive, the knowledge that biological networks can add another layer of information to deconvolve mixed-up omics data remains useful.

5.2 Synopsis

Motivation: Most microbes on Earth have never been grown in a laboratory, and can only be studied through DNA sequences. Environmental DNA sequence samples are complex mixtures of fragments from many different species, often unknown. There is a pressing need for methods that can reliably reconstruct genomes from complex metagenomic samples in order to address questions in ecology, bioremediation, and human health.

Results: We present the SOrting by NEtwork Completion (SONEC) approach for assigning reactions to incomplete metabolic networks based on a metabolite connectivity score. We successfully demonstrate proof of concept in a set of 100 genome-scale metabolic network reconstructions, and delineate the variables that impact reaction assignment accuracy. We further demonstrate

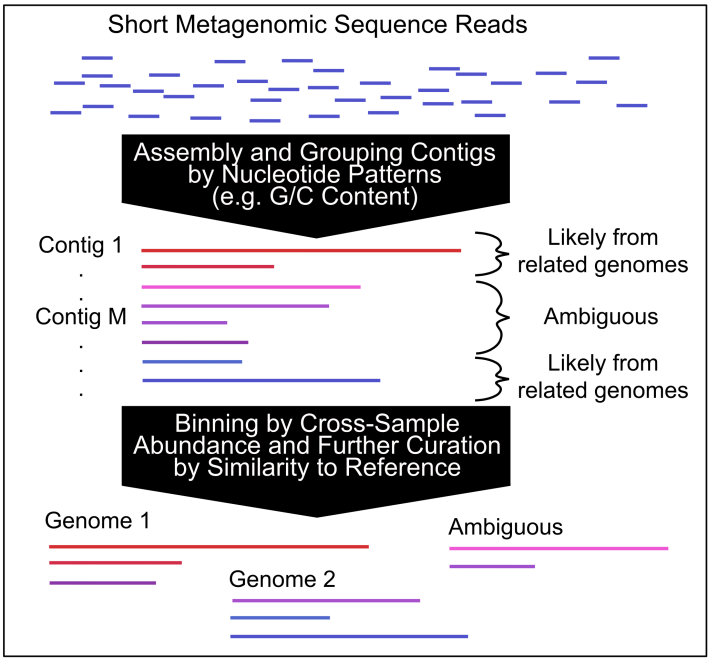


Figure 5.1: **Current approaches to reassembling species-level genomes from metagenomic data** include: assembly, where short reads are assembled into larger fragments ("contigs") by sequence overlap; grouping by sequence composition, where fragment similarity is gauged by nucleotide sequence patterns (e.g. G/C content or tetranucleotide frequencies); clustering by cross-sample abundance profiles, where fragments with strongly correlated abundance across independent samples are grouped together; further curation can include mapping to closely-related reference genomes, taxonomic annotations, or testing that bins contain minimal gene sets common to most organisms.

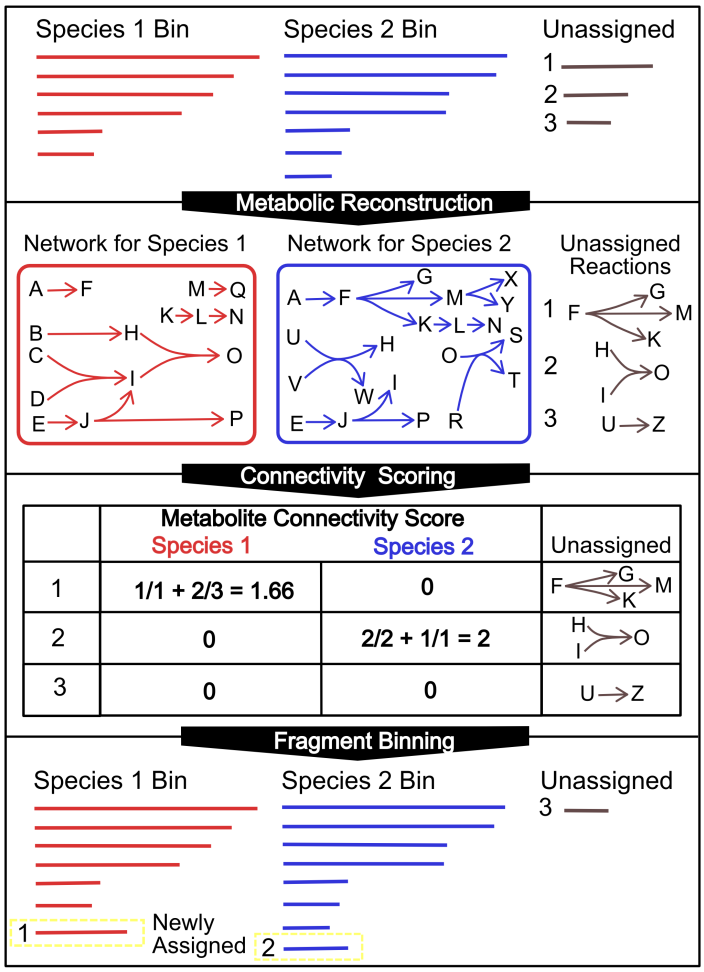


Figure 5.2: **The SONEC Algorithm**. The algorithm is initialized with bins of contigs (where the bins correspond to species from the metagenome), and a set of unassigned sequence fragments. A metabolic network reconstruction is produced for each bin and all unassigned contigs. To determine the correct parent bin to which unassigned contigs should be assigned, a metabolite connectivity score is calculated for each pair of unassigned reaction and parent network. This metabolite connectivity score quantifies the number of dead-end metabolites in the parent network which would no longer be dead-end with the addition of the unassigned reaction. Unassigned reactions will remove more dead-end metabolites, on average, from the correct parent network than from other, off-target networks. If there is a single maximum metabolite connectivity score for a given reaction, the contig associated with that reaction is assigned to the parent bin indicated by the metabolite connectivity score (e.g. unassigned contig 1 is assigned to species bin 1, and unassigned contig 2 is assigned to species bin 2, while unassigned contig 3 is ambiguous and cannot be assigned).

the integration of SONEC with existing approaches (such as cross-sample scaffold abundance profile clustering) on a set of 94 metagenomic samples from the Human Microbiome Project. We show that not only does SONEC aid in reconstructing species-level genomes, but it also improves functional predictions made with the resulting metabolic networks.

Availability and implementation: The datasets and code presented in this work are available at: bitbucket.org/mattbiggs/sorting_by_network_completion/.

5.3 Introduction

Most microbes cannot be cultured using existing techniques [1]. It is possible to interrogate this vast world of 'unculturables' by analysis of DNA from environmental samples. Metagenomics is a burgeoning field, and databases are accumulating trillions of bases of DNA sequence from complex environmental samples. These DNA fragments contain information about new and interesting microbes. Many approaches for analyzing such complex mixtures of DNA fragments seek to catalog the families of genes contained in the community metagenome, and how those families of genes change over time [2–4]. Other approaches seek to assign DNA fragments to known taxonomic groups [5, 6]. What is more difficult is the assignment of DNA fragments—genes in particular—to yet undiscovered parent genomes, and as a result, discovering the context in which those genes operate. The goal is not only to know that a given gene exists within the community, but to know also to which species that gene belongs, what other genes that species has, what metabolic capacity that species presents, the regulatory network that controls those genes and so on. Answers to these questions will advance efforts to discover new pathogens, industrially-relevant microbes and drivers of global geochemical cycles [7–9].

Recent advances in reconstructing species-level genomes from metagenomic samples have relied on several sources of information: nucleotide patterns that differentiate species, such as G/C content and tetranucleotide frequencies (Figure 5.1) [10, 11]; taxonomic assignment based on similar, known genomes [12]; improved fragment assembly [13]; and differential scaffold abundance across multiple samples [14–18]. The best approaches to-date use all of these sources of information to extract high quality, species- or strain-specific genomes [17]. While the best current approaches have demonstrated the ability to extract hundreds of genomes from a complex community such as the human gut, they still leave a third of the available DNA fragments unassigned [17].

We propose a new, orthogonal source of information that can be used to further improve species genome re-

construction, in conjunction with existing approaches. Metabolic networks are assumed to be effectively complete (i.e. gapless) [19–22]. This assumption of network completeness is a physiological equivalent of the law of conservation of mass: that is, that mass drawn into a cell must eventually leave or be integrated into biomass. Thus, real metabolic networks do not contain "dead end" metabolites—reaction substrates or products that are exclusively consumed or produced [19, 21]. This fact can be leveraged in the assignment of metagenomic fragments to species bins. Given a set of bins containing genetic fragments (formed using orthogonal sources of information as described above), and a set of unassigned fragments, a metabolic network can be reconstructed based on the gene content of each bin, and new fragments assigned to these bins based on a metabolite connectivity metric. The underlying assumption driving this approach is that genetic fragments containing metabolic genes will tend to fill gaps in the correct host metabolic network, and will be less likely to fill gaps in a foreign network to which they do not belong.

We refer to this approach as SOrting by NEtwork Completion (SONEC). We present proof-of-principle results from the successful application of this method using a set of 100 genome-scale metabolic network reconstructions. Furthermore, we demonstrate the application of this approach to 94 metagenomic samples from the Human Microbiome Project [23]. These computational experiments highlight the utility of this novel method, and delineate the sensitivity to variables that impact practical applications.

5.4 Methods

Obtaining metabolic network reconstructions

All metabolic network reconstructions were generated by the Model SEED server [24]. These were downloaded as spreadsheets and converted to Matlab objects using custom scripts, available in Supplementary Material [25]. To generate network reconstructions for each individual cluster in the anterior nares dataset, the set of 9910 assembled contigs was uploaded to the model SEED server. The reactions for each cluster were assigned by mapping the annotated open reading frames to the gene-protein-reaction associations in the meta-reconstruction. All 100 single-species reconstructions are publically available through the model SEED, and our copies of all reconstructions are also available as Matlab objects.

Proof-of-concept simulations

All simulations were performed using custom scripts in Matlab R2013a on a machine running 64-bit Windows 7, 32 GB RAM and 3.6GHz processor speed. Confidence intervals were calculated in R [26]. All scripts and data are available in the Supplementary Materials.

Binary error estimation

We organized errors into the following categories: True Positives (TP) result from the case where reactions were unambiguously assigned to the correct parent network; False Positives (FP) result from the case where reactions were unambiguously assigned to an incorrect network; True but Ambiguous (TA) results from the case where there were one or more ties in the maximum metabolite connectivity score (for definition of "metabolite connectivity score", see the "Algorithm" section of Results), and included the correct parent network; False and Ambiguous (FA) results from the case where there were one or more ties in the maximum metabolite connectivity score, none of which were the correct parent network; True Rejection (TR) results from the case where there was a metabolite connectivity score of zero for all networks and the rejected reaction originated from a shadow network (and thus, was correctly rejected; for definition of 'shadow network', see the 'Algorithm' section of Results); False Rejection (FR) results from the case where there was a metabolite connectivity score of zero for all networks, but the rejected reaction originated from one of the visible networks and so was incorrectly rejected. All error bars represent the 95% confidence interval for the observed accuracy. Because the assignments resulted in binary outputs (correct assignment or not), confidence intervals were estimated using the Wilson score interval [27] in R.

Obtaining metagenomic samples

Illumina whole-genome shotgun reads were obtained from the Human Microbiome Project database [23]. All 94 samples corresponding to the anterior nares were downloaded, while 49 samples containing more than one million reads were used to estimate coverage of assembled fragments. These 49 samples were each reduced to one million reads in order to normalize coverage estimates. This was done by randomly selecting one million reads from the total sample using a custom Python script (available in the Supplementary Material). The methods pertaining to the complete analysis of this metagenomic dataset can be found in the Supplementary Materials.

***In silico* reaction essentiality screen**

Reaction essentiality was determined by setting the upper and lower flux bounds to zero for each reaction in turn. Flux Balance Analysis was performed using the COBRA toolbox for Matlab [28] and the Gurobi Optimizer [29]. Reactions were considered essential if, when the reaction was prevented from carrying flux, flux through biomass was also reduced to zero. Visualization of the metabolic network and essential reactions was performed using MetDraw [30]. Our data and code are available in the Supplementary Material.

5.5 Results

Algorithm

We define a metabolite connectivity score (MCS) for reaction i with respect to metabolic network j as:

$$MCS_{ij} = \frac{|RS_i \cap NC_j|}{|RS_i|} + \frac{|RP_i \cap NP_j|}{|RP_i|}$$

where RS_i is the set of substrates for reaction i ($|RS_i|$ is the number of substrates for reaction i), RP_i is the set of products for reaction i, NC_j is the set of metabolites that are not consumed by any reaction in network j, and NP_j is the set of metabolites that are not produced by any reaction in network j. $\cap$ indicates the intersection between sets. Given an unassigned reaction and a set of metabolic networks, the metabolite connectivity score is calculated for each network and the reaction is assigned to the network with the maximum metabolite connectivity score (Figure 5.2). In the case of a tie, the correct assignment is ambiguous. In this work we chose to only assign reactions with unambiguous metabolite connectivity scores, but the algorithm could be readily adapted to make more liberal assignments.

Additionally, the concepts of "groups" and "shadow networks" are important for understanding the proof-of-concept simulations that follow. We define a "group" as a set of reactions that originate from the same metabolic network. A group can be thought of as a set of metabolic reactions that are obtained from genes on the same contiguous metagenomic sequence fragment (or "contig"), thus we can be confident that these genes come from the same parent organism. A group metabolite connectivity score is defined as the sum of the scores for each individual reaction: $MCS_{kj} = \sum_i^N MCS_{ij}$ where MCS_{kj} is the metabolite connectivity score for group k (of size N reactions) with respect to metabolic network j, and MCS_{ij} is the metabolite connectivity score for reaction i (within group k) with respect to metabolic network j.

We define "shadow networks" as a pool of metabolic networks which contribute reactions to the metagenome, but which are not considered as potential bins to which reactions can be assigned. For example, consider a metagenomic dataset with many high-abundance species and several low-abundance species. A bin can be created corresponding to each high-abundance species because there is sufficient signal in the dataset. However, species of very low-abundance in the community are probably not sequenced to sufficient depth to be assigned their own bins (in other words, this is a "shadow species"). Because the sequence fragments from these low-abundance species cannot be assigned to their own bins, they may be incorrectly assigned to bins of high-abundance species (because bins corresponding to high-abundance species are the only available choices for assignment). Including these "shadow networks" in our simulations allows us to evaluate the strength of the MCS in differentiating reactions that do not originate from any available choice of reconstruction.

While the analysis below demonstrates the value of SONEC, we provide here specific examples of binning based on the MCS to highlight the functionality and caveats of this scoring scheme (Supplemental Figure S1). Beginning with a set of 10 draft-quality metabolic network reconstructions, we randomly removed reactions from each and used the MCS to assign these reactions back to a metabolic network. As an example of a true positive result, the MCS was calculated for a reaction catalyzed by a 5-phosphomevalonate phosphotransferase drawn from *Enterococcus* sp. GMD1E (Supplemental Figure S1A). The metabolic network for *Enterococcus* sp. GMD1E was the only network of 10 that contained dead-end metabolites that overlapped with products of the reaction. In this case, diphosphomevalonate was not produced by any reaction in the *Enterococcus* network, and the MCS captured this complementary overlap with the reaction products, resulting in a correct assignment.

In contrast, an example of a false positive result is informative (Supplemental Figure S1B). The reaction catalyzed by a nicotinate-nucleotide dimethylbenzimidazole phosphoribosyltransferase overlapped with dead-end metabolites in several networks. In the correct parent network of *Shigella flexneri*, the reaction product—alpha-ribazole 5′-phosphate—was an unproduced metabolite. Because there were three products in the reaction, the MCS is 0.33. Conversely, in the network for Pelagibacter ubique, a reaction substrate nicotinate ribonucleotide was an unconsumed metabolite. Because there are only two substrates in the reaction, the MCS is 0.5, and because this was the maximum, the reaction was incorrectly assigned to *P. ubique*. These specific examples of true and false positives exhibit how

the MCS works in practice. We performed further simulations which help to elucidate the role of variables that influence reaction assignment accuracy using the MCS.

Proof-of-concept simulations

We simulated the problem of binning metagenomic samples into appropriate species bins. For each independent simulation, we started with a set of draft-quality metabolic network reconstructions randomly drawn from among 100 bacterial networks obtained from the Model SEED. We randomly removed reactions from each. We used the MCS to assign these reactions back to a metabolic network reconstruction. Unless otherwise noted, accuracy of assignment was evaluated over 1000 independent simulations for every unique combination of parameters. While the simulations presented here are performed with networks from the Model SEED, the same analysis could be performed with networks derived from other resources such as KEGG or Pathway Tools [31, 32], ideally with more coverage of known microbial taxa.

We first evaluated the effect of parent network completeness on reaction assignment accuracy (Figure 5.3A). We randomly removed increasingly large subsets of reactions from each of 10 metabolic networks. These reactions were then assigned to a network. More complete parent networks (fewer reactions removed from the original) produced more true positive, and fewer false positive, reaction assignments. As expected, as parent networks become less complete, assignment accuracy diminishes with decreases in true positives and increases in false positives. Each simulation was repeated 1000 times, with a new set of 10 parent networks being selected randomly each time from the pool of 100 networks. We display results from group sizes (number of reactions annotated from the same sequence fragment or contig) of 1, 20 and 40 (Figure 5.3A).

Next, we investigated the impact of increasing the number of parent networks from which unassigned reactions were derived (Figure 5.3B). For these simulations, the fraction of reactions removed was fixed at 0.15 and the group size fixed at 25. As the number of networks increased, we observed corresponding decreases in the number of true positives and increases in the number of false positive reaction assignments.

We further evaluated the effect of increasing group size (Figure 5.3C). The fraction of reactions removed was fixed at 0.15, and the number of parent networks was fixed at 10. We observed that increasing the group size improved assignment accuracy. Group sizes less than five tended to produce true but ambiguous assignments. Group sizes of six or greater produced mostly true positive assignments, with a steadily improving margin between true and false positives as group size increased. The interaction between network completeness and group size (or any other combination of parameters) can be evaluated extensively through further simulations (Supplemental Figure S2).

We also explored the impact of shadow networks (networks which contribute reactions to the unassigned pool, but do not have a corresponding bin to which reactions can be assigned) (Figure 5.3D). The fraction of reactions removed was fixed at 0.15, the number of parent networks fixed at 10 and the group size fixed at 25. Reactions drawn from shadow networks were included for assignment, but the shadow networks were not included as candidates to which reactions could be assigned. We observed an interesting pattern of assignment accuracy as the number of shadow networks increased (displayed as a fraction of the total population of networks). True positive assignments account for the majority, up until the number of shadow networks is equivalent $\sim$0.2 of the population. Between 0.23 and 0.44, false positives account for the majority and from 0.44 to 0.5, true but ambiguous assignments form the majority. We also observed an interesting increase in false and ambiguous assignments that peaked at $\sim$0.3.

Finally, we evaluated the impact of SONEC on functional network predictions by comparing reaction essentiality predictions from pre- and post-SONEC networks to the predictions from the full, parent network (Figure 5.3E). In this set of simulations, the fraction of reactions removed was fixed at 0.15, the number of parent networks fixed at 10 and the group size fixed at 25, over 50 replicates. In each replicate, one network was chosen for evaluation. After reactions were removed, a copy of the incomplete network (the pre-SONEC network) was gap-filled [33]. Subsequently, all reactions assigned by SONEC were added to a separate copy of the imcomplete network (the post-SONEC network), which was then gap-filled. Reaction essentiality for the pre-SONEC, post-SONEC and full networks were all evaluated using the same biomass function and exchange flux bounds. The post-SONEC reaction essentiality predictions achieved accuracies 1.8% greater, on average, than the pre-SONEC predictions (p-value=$2.7x10^6$ by paired, one-sided Wilcoxon rank sum test). The post-SONEC predictions were the same or better 80% of the time, and strictly better 68% of the time. In this case, we evaluated reaction essentiality rather than gene essentiality (a more common measure) due to the draft-quality status of the gene-protein-reaction relationships.

Example values of SONEC parameters (e.g. group size, network completeness) in existing metagenomic datasets are described in the supplemental materials.

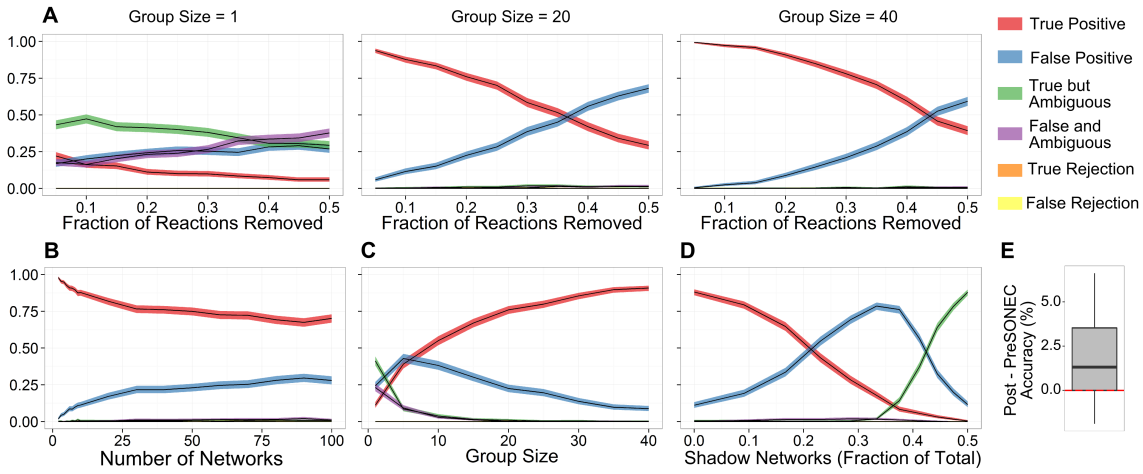


Figure 5.3: **Reaction assignment accuracy from simulations of the SONEC approach**. (A) The accuracy is displayed as a function of the number of reactions that were removed from the parent networks (shown as a fraction of parent network size). An increasing number of reactions were removed from the total reaction content of 10 randomly-selected metabolic networks. Results for group sizes (the number of reactions being assigned together) of 1, 20 and 40 are shown. (B) Accuracy is displayed as a function of the number of parent bins to which a reaction could potentially be assigned. The fraction of reactions removed was fixed at 0.15, and the group size fixed at 25. (C) Accuracy is displayed as a function of reaction group size. The fraction of reactions removed was fixed at 0.15, and the number of parent networks was fixed at 10. (D) Accuracy is displayed as a function of the number of shadow networks. Shadow networks are a pool of network reconstructions from which reactions are contributed to the unassigned pool, but which are not available as bins to which those reactions can be assigned. The fraction of reactions removed was fixed at 0.15, the number of visible parent networks fixed at 10 and the group size fixed at 25. (E) Reaction essentiality predictions from gap-filled networks pre- and post-SONEC were compared to predictions from the full, reference networks. The difference in accuracy between paired experiments (post-SONEC - pre-SONEC) is shown here as a boxplot, with the null hypothesis (zero, no difference) indicated by the dashed, red line. SONEC improved the average accuracy by 1.8%, with a p-value of 2.7×10^6 (by paired, one-sided Wilcoxon rank sum test on 50 replicates). For A–D, all results are from 1,000 independent replicates and shaded areas represent a 95% confidence interval around the mean, determined by the Wilson score interval (see Section 2).

Pathway enrichment

Pathway enrichment was performed to evaluate the contribution of different families of metabolic reactions to assignment accuracy (Supplemental Figure 3 and Supplemental Methods). For these simulations, 10 parent networks were available for assignment, the fraction of reactions removed was fixed at 0.15, the group size fixed at 25, and there were no reactions from shadow networks. Sulfur metabolism, nicotinate and nicotinamide metabolism, galactose metabolism, porphyrin and chlorophyll metabolism, biosynthesis of steroids, and terpenoid biosynthesis contributed to true positive assignments more than expected by chance alone. Propanoate metabolism, pyruvate metabolism, amino sugars metabolism, and others contributed to more false positive assignments than expected by chance alone. Several pathways, including ubiquinone biosynthesis, D-glutamine and D-glutamate metabolism, were all enriched in both true positive and false positive assignments.

SONEC applied to metagenomic samples

We applied the SONEC approach to metagenomic sequences from 94 samples sourced from the human anterior nares as part of the Human Microbiome Project (Figure 5.4) [23]. The short reads were assembled into contigs, and the abundance of each contig was estimated across the subset of 49 samples containing more than one million reads. The assembly process resulted in 1,543,959 contigs, with an N50 of 261. The N50 indicates the contig length at which all contigs of that length or greater contribute 50% of the cumulative length of the dataset. We continued the analysis with the 9,910 contigs with length of 800 base pairs (bp) or greater and which had a non-zero abundance in at least 3 samples (Figure 5.4A). These contigs were clustered into 2,849 clusters, such that contigs within a cluster likely originated from the same organism (Figure 5.4B). Metabolic network reconstructions were obtained for all clusters by uploading the corresponding contigs to the Model SEED server resulting in 14 083 annotations including

open reading frames and RNA elements [24]. We observed 14 clusters with 90 or more annotated reactions and an average cumulative length of 682,232bp. The remaining smaller clusters contained 44 or fewer annotated reactions and an average cumulative length of 2171bp. The taxonomic content of each cluster was estimated, and the two large clusters with the most consistent taxonomic identity (Clusters 614 and 1,357) corresponded to strains of *Staphylococcus aureus*. Cluster 614 (labeled "Strain 1") contained 1,001 assembled fragments with a cumulative length of 1,623,468bp, 742 assigned metabolic reactions, and 100% of fragments aligned well to S.aureus genomes. Cluster 1357 (labeled "Strain 2") contained 396 assembled fragments with a cumulative length of 479,515bp, 326 assigned metabolic reactions, and 90% of fragments aligned well to *S.aureus* genomes. For reference, the complete genome for *S.aureus* Newman is 2.9 million bp long [34]. These two clusters were not correlated and thus, likely originated from different strains of *S.aureus* (Figure 5.4A).

Comparing Strain 1 to a reference metabolic network for *S.aureus* N315 (obtained from the Model SEED) revealed 692 shared reactions of a possible 1,118. Strain 1 contained 50 unique reactions that were not found in the reference metabolic network. These unique reactions were found in the following pathways: biosynthesis of steroids; butanoate metabolism; glycine, serine and threonine metabolism; pentose and glucuronate interconversions; pentose phosphate pathway. Strain 2 shared 319 reactions with the *S.aureus* N315 reference. Strain 2 contained a further 7 unique reactions in the following pathways: glutathione metabolism; pentose phosphate pathway; purine metabolism; pyrimidine metabolism; pyruvate metabolism.

After using established techniques [14, 17] to identify Strain 1 and Strain 2, we applied SONEC to further complete these two clusters. Seven smaller clusters were identified as also originating from strains of *S.aureus*. These smaller clusters had cumulative lengths from 920 to 59378bp, were annotated with 10–42 metabolic reactions, and 100% of fragments aligned well to *S.aureus* genomes. The SONEC MCS was utilized to assign these smaller clusters to one of the larger *S.aureus* clusters. Five of the seven clusters produced non-zero metabolite connectivity scores and could be assigned unambiguously. Two were assigned to Strain 1 and three to Strain 2, which increased reaction overlap with the reference metabolic network for *S.aureus* N315 by 3.8% and 7.8% respectively. Many of the newly assigned reactions expanded core subsystems such as glycolysis and amino acid metabolism (Figure 5.4C). The addition of these smaller clusters increased the total genetic content of Strain 1 by 2,968bp and Strain 2 by 69,913bp. To determine the impact of these SONEC assignments on functional predictions of the resulting metabolic networks, we performed an *in silico* reaction essentiality screen on Strain 1 before and after the application of SONEC (Figure 5.5). To begin, we identified an *S.aureus* minimal medium and a biomass function [35]. We performed gap filling based on the identified medium and biomass formulation using a custom implementation of a previously described gap fill algorithm (code available in the Supplementary Material) [33]. We chose candidate reactions for gap filling from the complete Model SEED reaction database [24]. We gap filled Strain 1 before and after the application of SONEC (Figure 5.5A), and evaluated the essentiality of all reactions (excluding the reactions added during the gap filling process). There were 14 reactions which were essential before SONEC, but not after (Figure 5.5B). An example of these is a nucleosidase classified under methionine metabolism. There were 18 reactions which became essential after SONEC but were not essential beforehand. An example of these is an oxidoreductase found in glutamate and arginine metabolic pathways. There were 14 reactions which were constitutively essential. Interestingly, no reaction added by SONEC was essential.

5.6 Discussion

Here we present the SONEC approach for the assignment of metabolic reactions (and as an extension, metagenomic sequence fragments annotated with metabolic genes) back to a parent metabolic network. This work is motivated by the fact that current approaches are still unable to group complete metagenomic samples into member genomes, leaving, in a recent study, 32% of metagenomic sequence fragments unaccounted for [17].

We propose that information about the metabolic network can be used to improve metagenomic fragment binning. It is commonly assumed that metabolic networks are gapless, and gap filling of metabolic network reconstructions is used regularly as a source of new biological knowledge [19, 21, 33, 36]. Here, we demonstrate that gap filling can similarly be used to assign reactions to the correct parent metabolic network by using a metabolite connectivity score, and thus improve metagenome sequence annotation (Figure 5.2).

We observe that more complete networks (reconstructed from bins of metagenomic sequence fragments) initially lead to improved reaction assignment accuracy (Figure 5.3A). As parent networks degrade and lose more and more reaction content, accuracy is lost. This observation aligns with intuition, as more complete networks provide context in which to place new reactions.

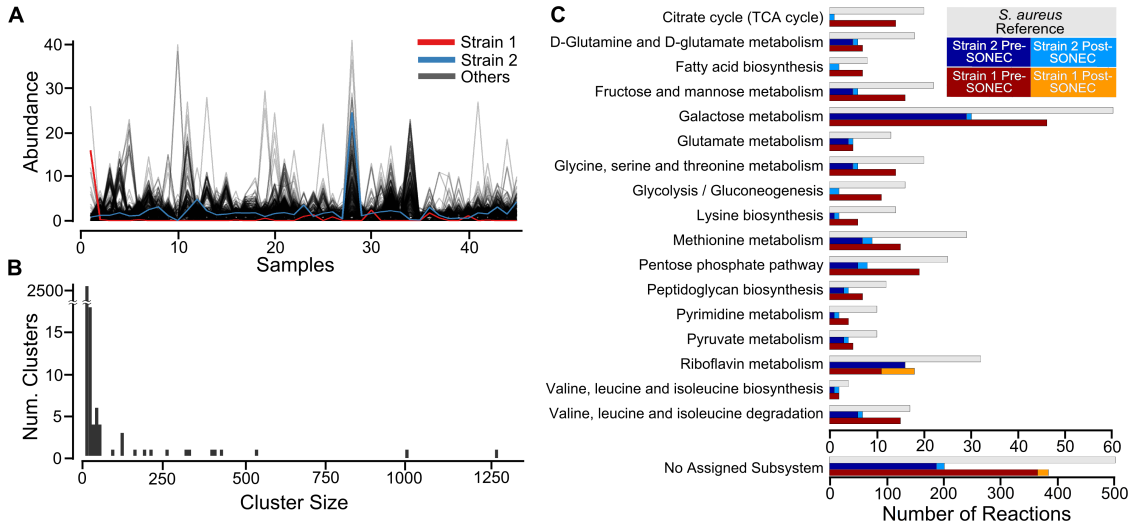


Figure 5.4: **Generating species-level clusters from the anterior nares metagenome data set**. (A) Average cross-sample abundance profiles for all 2,849 clusters after application of the canopy algorithm. The profiles for the two clusters which we refer to as Strain 1 and Strain 2 are highlighted in red and blue, respectively. (B) Histogram of cluster size (number of contigs) for all clusters after application of the canopy algorithm [17]. Note that most clusters are very small (2,500 clusters with fewer than 10 contigs), while there are few very large clusters. (C) Reaction content of metabolic network reconstructions, organized by subsystem, for Strain 1 and Strain 2, before and after the application of SONEC. Reaction content from a reference network for *S.aureus* is provided.

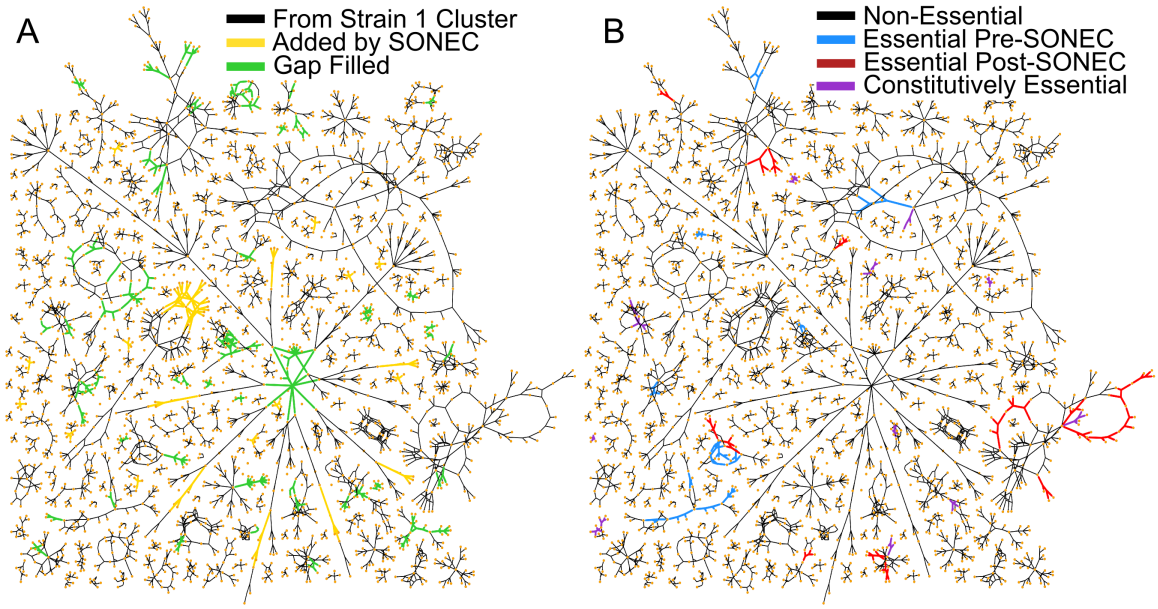


Figure 5.5: **Application of SONEC alters functional predictions of metabolic network**. In both panels the metabolic network for Strain 1 is represented with metabolites as nodes (orange circles) and reactions as edges between metabolites. (A) Reactions are colored by source: black indicates reactions from the original cluster for Strain 1; yellow indicates reactions added by SONEC; and green reactions were added during the gap filling process. (B) Reactions are colored to indicate essentiality: black reactions are non-essential in all conditions; blue reactions were essential before the application of SONEC, but not after; red reactions were essential after the application of SONEC, but not before; and purple indicates reactions which were essential before and after SONEC.

Similarly, as the number of parent networks increases, accuracy is lost (Figure 5.3B). This observation also makes sense, recognizing that the presence of more networks increases the opportunity to mis-assign a reaction. Encouragingly, increasing group (a set of metabolic reactions known to originate from the same organism) size significantly improves reaction assignment accuracy under all conditions (Figure 5.3C). Group size can be in-

creased by improved assembly or fragment clustering—anything that will increase the number of genes that can be confidently associated with each other. While any given reaction may fill gaps in several possible networks, the likelihood is low of an entire group of reactions filling gaps in the same, incorrect, network. In other words, for large groups of reactions, the error is diluted over the many possible wrong choices, while the metabolite connectivity score accrues for the correct parent network. The presence of shadow networks takes a toll on accuracy (Figure 5.3D). Shadow networks can be thought of as the set of organisms in the community that contributed metagenomic sequence, but were not assigned bins. Therefore, any attempt to assign those reactions to existing bins will be incorrect. Finally, simulations showed that functional network predictions (Figure 5.3E) are generally improved by SONEC, an outcome that has significant implications for application of subsequent metabolic network analysis (Figure 5.5). Interestingly, none of the reactions added by SONEC were essential. However, by adding them, the network structure changed in such a way as to make some previously essential reactions non-essential, and vice versa. One possible explanation for this improvement is that SONEC assigns reactions in a relatively unbiased way (based on metabolite connectivity) compared to traditional gap-filling, which adds reactions to allow flux through a biomass function. Future applications which require functional predictions of the impact of genome engineering or drug targeting within microbial communities can benefit from SONEC. In the end, the goal of SONEC is to improve the reconstruction of individual genomes from metagenomic data. More complete genomes will improve any downstream analyses.

Future work may improve assignment accuracy by modifying the metabolite connectivity score. The example in Supplemental Figure S1 highlights a weakness of the metabolite connectivity score, wherein two models may contain a single dead-end metabolite that overlaps with a reaction, but depending on whether it is a substrate or product, the final gap score may be different. Maintaining the ratios in the metabolite connectivity score is prudent from a parsimony standpoint, because they ensure that the smallest reaction (with the fewest participating metabolites) that can fill a gap will be used. However, future work could explore alternative metabolite connectivity scores that address the weaknesses with the scoring framework presented here. One possibility would be to penalize the addition of new metabolites that do not exist in the network, which would have improved the outcome for the false positive example in Supplemental Figure S1. This may prohibit filling larger gaps consisting of more than one reaction, or filling gaps in less complete networks. Another approach is to apply a global optimization-based gap fill algorithm based on existing methods [33]. We chose not to pursue this approach because it would be sensitive to the choice of optimization function and exchange constraints, which are difficult to determine for uncharacterized microbes in complex environments.

Enrichment analysis highlights the families of reactions that tend to provide better assignment accuracy (Supplemental Figure S3). The underlying driver may be that reactions that contain uncommon metabolites are more likely to be assigned to the correct parent network. Within the selection of 100 prokaryotic reconstructions used here, porphyrin and chlorophyll metabolism are uncommon. Given this hypothesis, future work may improve assignment accuracy by selectively weighting reactions that are unique within the environment being studied. For example, in the anterior nares dataset explored here, the rarest pathways include lipoic acid metabolism, inositol metabolism and caprolactam degradation. To improve group assignment accuracy, metabolite connectivity scores corresponding to reactions from these rare subsystems would be weighted more heavily (as they would be expected to increase accuracy disproportionately).

To demonstrate how the SONEC approach can be applied to real metagenomic data, we analyzed 94 metagenomic samples sourced from the human anterior nares (Figure 5.4). It is important to note that these samples were not sequenced very deeply, and as a result, the N50 we could achieve after assembly was quite low (250bp). As a comparison, a recent study assembled DNA fragments from stool samples to achieve an N50 of more than 40,000bp [37]. This observation simply indicates that in applications with deeper sequencing, contigs will tend to be much longer. Knowing that larger group size—which is a function of longer contigs—improves SONEC performance, it is likely that SONEC performance will improve with deeper sequencing and more complete assembly. While it is clear that the performance of SONEC is highly dependent on the existing tools used to create the initial bins, the simulations we performed demonstrate that SONEC can add value and improve predictions even with imperfect data.

We first applied established approaches to create initial clusters of metagenomic sequence fragments, including short read assembly and clustering by cross-sample abundance and nucleotide composition patterns. A BLAST-based estimate of cluster taxonomic consistency (that is, the percentage of fragments within the cluster that map to the same taxonomy) revealed that of the large clusters, only two clusters were >90% consistent. This consistency can be compared to a larger-scale

study which analyzed 396 microbiome samples from the human gut, in which 115 large clusters were found to be >95% consistent [17]. Clearly, it is possible to improve the initial clustering and conditions before applying SONEC. Given the two large clusters which mapped consistently to strains of *S.aureus*, we demonstrated how SONEC can be used to assign smaller, orphan clusters to these larger clusters. This practical demonstration on real data shows that by including metabolic information, ambiguous fragments can be assigned to the parent genomes. As a quality check, the resulting metabolic networks after applying SONEC are more consistent with a reference S.aureus metabolic network reconstruction.

5.7 Acknowledgments

This work was supported by the National Institutes of Health [grant number R01 GM108501 to JP], a Jefferson Trust Big Data Fellowship [to MBB], and a National Institutes of Health Training Grant [project number 2T32GM008715-16] through the University of Virginia [to MBB].

The authors would like to thank Phillip Yen for his help utilizing the computational resources at UVA.

5.8 References

[1] Handelsman J. "Metagenomics: application of genomics to uncultured microorganisms." In: *Microbiology and molecular biology reviews* 68.4 (2004), pp. 669–685. DOI: 10.1128/MMBR.68.4.669-685.2004.

[2] Abubucker S et al. "Metabolic reconstruction for metagenomic data and its application to the human microbiome". In: *PLoS Computational Biology* 8 (2012). DOI: 10.1371/journal.pcbi.1002358.

[3] Greenblum S, Turnbaugh PJ, and Borenstein E. "Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease". In: *Proceedings of the National Academy of Sciences* 109 (2012), pp. 594–599. DOI: 10.1073/pnas.1116053109/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1116053109.

[4] Owen JG et al. "Multiplexed metagenome mining using short DNA sequence tags facilitates targeted discovery of epoxyketone proteasome inhibitors". In: *Proceedings of the National Academy of Sciences* (2015), p. 201501124. DOI: 10.1073/pnas.1501124112.

[5] Afshinnekoo E et al. "Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics". In: *Cell Systems* (2015). DOI: 10.1016/j.cels.2015.01.001.

[6] Yarza P et al. "Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences". In: *Nature Reviews. Microbiology* 12 (2014), pp. 635–645. DOI: 10.1038/nrmicro3330.

[7] Kinross JM, Darzi AW, and Nicholson JK. "Gut microbiome-host interactions in health and disease." In: *Genome medicine* 3 (2011), p. 14. DOI: 10.1186/gm228.

[8] Rousk J and Bengtson P. "Microbial regulation of global biogeochemical cycles". In: *Frontiers in Microbiology* 5.March (2014), pp. 305–7. DOI: 10.3389/fmicb.2014.00103.

[9] Smid EJ et al. "Functional implications of the microbial community structure of undefined mesophilic starter cultures". In: *Microbial Cell Factories* 13.Suppl 1 (2014), S2. DOI: 10.1186/1475-2859-13-S1-S2.

[10] Iverson V et al. *Untangling Genomes from Metagenomes: Revealing an Uncultured Class of Marine Euryarchaeota*. 2012. DOI: 10.1126/science.1212665.

[11] Teeling H et al. "Application of tetranucleotide frequencies for the assignment of genomic fragments." In: *Environmental microbiology* 6.9 (Sept. 2004), pp. 938–47. DOI: 10.1111/j.1462-2920.2004.00624.x.

[12] MacDonald NJ, Parks DH, and Beiko RG. "Rapid identification of high-confidence taxonomic assignments for metagenomic data". In: *Nucleic Acids Research* 40 (2012). DOI: 10.1093/nar/gks335.

[13] Namiki T, Hachiya T, Tanaka H, and Sakakibara Y. "MetaVelvet: An extension of Velvet assembler to de novo metagenome assembly from short sequence reads". In: *Nucleic Acids Research* 40 (2012), pp. 1–12. DOI: 10.1093/nar/gks678.

[14] Albertsen M et al. "Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes." In: *Nature biotechnology* 31.6 (2013), pp. 533–8. DOI: 10.1038/nbt.2579.

[15] Alneberg J et al. "Binning metagenomic contigs by coverage and composition". In: *Nature Methods* (2014). DOI: 10.1038/nmeth.3103.

[16] Carr R, Shen-Orr SS, and Borenstein E. "Reconstructing the Genomic Content of Microbiome Taxa through Shotgun Metagenomic Deconvolution". In: *PLoS Computational Biology* 9 (2013). DOI: 10.1371/journal.pcbi.1003292.

[17] Nielsen HB et al. "Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes." In: *Nature biotechnology* 32.8 (2014), pp. 822–828. DOI: 10.1038/nbt.2939.

[18] Sharon I et al. "Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization". In: *Genome Research* 23 (2013), pp. 111–120. DOI: 10.1101/gr.142315.112.

[19] Krumholz EW and Libourel IGL. "Sequence-Based Network Completion Reveals the Integrality of Missing Reactions in Metabolic Networks". In: *Journal of Biological Chemistry* (2015), jbc.M114.634121. DOI: 10.1074/jbc.M114.634121.

[20] Pitkänen E et al. "Comparative genome-scale reconstruction of gapless metabolic networks for present and ancestral species." In: *PLoS computational biology* 10.2 (2014), e1003465. DOI: 10.1371/journal.pcbi.1003465.

[21] Satish Kumar V, Dasika MS, and Maranas CD. "Optimization based automated curation of metabolic reconstructions." In: *BMC bioinformatics* 8 (2007), p. 212. DOI: 10.1186/1471-2105-8-212.

[22] Thiele I and Palsson BØ. "A protocol for generating a high-quality genome-scale metabolic reconstruction." In: *Nature protocols* 5.1 (2010), pp. 93–121. DOI: 10.1038/nprot.2009.203.

[23] Huttenhower C et al. *Structure, function and diversity of the healthy human microbiome*. 2012. DOI: 10.1038/nature11234.

[24] Overbeek R. "The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes". In: *Nucleic Acids Research* 33.17 (2005), pp. 5691–5702. DOI: 10.1093/nar/gki866.

[25] MathWorks. *MATLAB and Statistics Toolbox*. Natick, Massachusetts, USA, 2012.

[26] R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria, 2016.

[27] Agresti A and Coull BA. "Approximate Is Better than "Exact" for Interval Estimation of Binomial Proportions". In: *The American Statistician* 52 (1998), pp. 119–126. DOI: 10.2307/2685469.

[28] Schellenberger J et al. "Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0." In: *Nature protocols* 6 (2011), pp. 1290–1307. DOI: 10.1038/nprot.2011.308.

[29] Gurobi. *Gurobi Optimizer*. Houston, TX, 2013.

[30] Jensen PA and Papin JA. "MetDraw: Automated visualization of genome-scale metabolic network reconstructions and high-throughput data". In: *Bioinformatics* 30 (2014), pp. 1327–1328. DOI: 10.1093/bioinformatics/btt758.

[31] Kanehisa M et al. "KEGG for representation and analysis of molecular networks involving diseases and drugs". In: *Nucleic Acids Research* 38.Database (Jan. 2010), pp. D355–D360. DOI: 10.1093/nar/gkp896.

[32] Karp PD et al. "Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology". In: *Briefings in Bioinformatics* 11.1 (Jan. 2010), pp. 40–79. DOI: 10.1093/bib/bbp043.

[33] Reed JL et al. "Systems approach to refining genome annotation." In: *Proceedings of the National Academy of Sciences of the United States of America* 103.46 (Nov. 2006), pp. 17480–4. DOI: 10.1073/pnas.0603364103.

[34] Baba T et al. "Genome sequence of Staphylococcus aureus strain newman and comparative analysis of staphylococcal genomes: Polymorphism and evolution of two major pathogenicity islands". In: *Journal of Bacteriology* 190 (2008), pp. 300–310. DOI: 10.1128/JB.01000-07.

[35] Becker SA and Palsson BØ. "Genome-scale reconstruction of the metabolic network in Staphylococcus aureus N315: an initial draft to the two-dimensional annotation." In: *BMC microbiology* 5 (2005), p. 8. DOI: 10.1186/1471-2180-5-8.

[36] Bartell JA et al. "Comparative Metabolic Systems Analysis of Pathogenic Burkholderia". In: *Journal of Bacteriology* 196.2 (2014), pp. 210–226. DOI: 10.1128/JB.00997-13.

[37] Jeraldo P et al. "Draft Genome Sequences of 24 Microbial Strains Assembled from Direct Sequencing from 4 Stool Samples: TABLE 1". In: *Genome Announcements* 3 (2015), e00526–15. DOI: 10.1128/genomeA.00526-15.

Chapter 6

Managing Uncertainty in Metabolic Network Structure and Improving Predictions Using EnsembleFBA

6.1 Context

As the last computational project completed during my graduate work, this project was the most organized, best-documented, most reproducible and most computationally sophisticated. At the beginning of this project, I grew immensely through long wrestles with linear algebra and optimization theory. This project was also my first to be shared on the preprint server bioRxiv (biorxiv.org) which, especially as a senior graduate student, made it very convenient to immediately share my most recent work. At the time of writing this dissertation, this work was under review at *PLOS Computational Biology*.

6.2 Synopsis

Genome-scale metabolic network reconstructions (GENREs) are repositories of knowledge about the metabolic processes that occur in an organism. GENREs have been used to discover and interpret metabolic functions, and to engineer novel network structures. A major barrier preventing more widespread use of GENREs, particularly to study non-model organisms, is the extensive time required to produce a high-quality GENRE. Many automated approaches have been developed which reduce this time requirement, but automatically-reconstructed draft GENREs still require curation before useful predictions can be made. We present a novel ensemble approach to the analysis of GENREs which improves the predictive capabilities of draft GENREs and is compatible with many automated reconstruction approaches. We refer to this new approach as Ensemble Flux Balance Analysis (EnsembleFBA). We validate EnsembleFBA by predicting growth and gene essentiality in the model organism *Pseudomonas aeruginosa* UCBPP-PA14. We demonstrate how EnsembleFBA can be included in a systems biology workflow by predicting essential genes in six *Streptococcus* species and mapping the essential genes to small molecule ligands from DrugBank. We found that some metabolic subsystems contribute disproportionately to the set of predicted essential reactions in a way that is unique to each *Streptococcus* species. These species-specific network structures lead to species-specific outcomes from small molecule interactions. Through these analyses of *P. aeruginosa* and six *Streptococci*, we show that ensembles increase the quality of predictions without drastically increasing reconstruction time, thus making GENRE approaches more practical for applications which require predictions for many non-model organisms. All of our functions and accompanying example code are available in an open online repository.

6.3 Introduction

Metabolism is the driving force behind the wondrous flurry of biological activity carpeting our planet. An organism's metabolism is determined by the metabolic enzymes encoded in its genome, the chemical reactions catalyzed by those enzymes, and whether or not those enzymes are actively expressed [1]. The simplest bacteria have hundreds of metabolic enzymes, while the most complex eukaryotes have thousands. The products of these enzymatic reactions serve as substrates for other reactions, such that the chemical transformations carried out in a cell can be represented as a vast network [2]. Mass and energy flow through such networks, transforming environmental inputs into the building blocks of life. Every species has a unique metabolic network driving its growth and interaction with the environment.

Genome-scale metabolic network reconstructions (GENREs) are formal representations of metabolic networks [3]. GENREs serve as a comprehensive collection of metabolic knowledge about a particular organism and they are amenable to mathematical analysis [4]. The process of reconstructing a GENRE takes months to years, but the reconstruction process often leads to new discoveries [5]. Mathematical analysis of GENREs gives insight into how particular metabolic pathways are used by an organism, what substrates it can utilize, which of its genes are essential in a given environment, how a

metabolic network can be engineered to produce more of a desired product, or which enzymes within the network should be targeted in order to halt growth in an organism [6–8]. The reconstruction and analysis of GENREs for single species has greatly contributed to our understanding of microbes and our ability to engineer them. Recently, analyses have been developed which predict metabolic interactions between microbes [9, 10]. However, the application of these recent analyses has been greatly limited by the large investment in time required to reconstruct a useful GENRE. Many microbial communities of interest consist of hundreds of species [11, 12]. It is decidedly impractical to spend decades manually curating hundreds of GENREs.

Many automated methods have been developed for rapidly reconstructing more accurate GENREs [13–16]. We present a novel ensemble method that is complimentary to these existing automated methods which we refer to as Ensemble Flux Balance Analysis (EnsembleFBA). EnsembleFBA pools predictions from many draft GENREs in order to more reliably predict properties that arise from metabolic network structure, such as nutrient utilization and gene essentiality (Figure 6.1). The primary benefits of this new method are that it relies on automatically-generated GENREs (which can be generated in a matter of minutes to hours) and yet produces more reliable predictions than individual GENREs within the ensemble. We implement and discuss one possible way of generating useful ensembles, but emphasize that other automated methods could be modified to generate useful ensembles.

We begin by discussing a common GENRE curation procedure known as gap filling. We demonstrate that a global gap filling procedure does not perform any better than a sequential one. Instead, we introduce an ensemble approach to pool the many possible network structures resulting from different sequences of the input media conditions (Figure 6.1). We demonstrate that an ensemble reliably outperforms most of its constituent GENREs in terms of predicting growth and gene essentiality. By tuning the stringency of the voting threshold (e.g. requiring a majority of GENREs to agree vs. complete consensus) it is possible to achieve greater precision or recall than any of the constituent GENREs. We show how additional steps to increase the diversity among GENREs within the ensemble (e.g. reconstructing each member GENRE using subsets of the available data) can further improve recall. Furthermore, we found that incorporating negative growth information into our GENREs improved overall accuracy of the ensemble. We present proof of concept of the use of ensembles by predicting carbon source utilization and gene essentiality in *Pseudomonas aeruginosa*, a well-studied, clinically-relevant pathogen. We provide an example workflow using EnsembleFBA by predicting gene essentiality in six *Streptococcus* species and mapping the predicted essential genes to small molecules ligands in DrugBank. All of our data and code are available in an online repository, including example scripts to make adoption of EnsembleFBA easy. Our ability to make mechanistic predictions about complex cellular communities requires advances in the way we leverage the data available to us, and the way we handle uncertainty. Ensemble FBA is a novel tool that maintains the speed of automated reconstruction methods while improving predictions by intentionally managing uncertainty in network structures.

6.4 Results

Gap Filling Against Multiple Media Conditions in Different Orders Produces Different Network Structures

Gap filling is the process of identifying mismatches between computational predictions and experimental results, and identifying changes to the network structure which will bring the computational predictions into agreement with the experimental data. "Gaps" are missing reactions and can be filled by drawing from a database of possible metabolic reactions. Given that there are usually many mismatches between computational and experimental results, we demonstrate that simply changing the order in which computational results are brought into agreement with experimental can result in different network structures. For example, suppose that it is experimentally determined that a microbe can grow on glucose minimal media and sucrose minimal media, but the computational predictions do not match. Gap filling the GENRE against a representation of glucose minimal media first and sucrose minimal media second, may result in a different network in the end than if sucrose minimal media were first. In practice, the order of gap filling is arbitrary.

We implemented a custom gap fill algorithm based on the algorithms FASTGAPFILL and FastGapFilling (see Materials and Methods) [17, 18]. We used the Model SEED biochemistry database as our "universal" reaction database from which to draw reactions for gap filling [13]. We used the Model SEED web interface to automatically generate a draft GENRE for *Pseudomonas aeruginosa* UCBPP-PA14 (without using the Model SEED gap filling feature). We gap filled this draft GENRE using 2, 5, 10, 15, 20, 25 and 30 media conditions that experimentally support growth, with 30 replicates for each. For example, we selected five media conditions at random and selected two random permu-

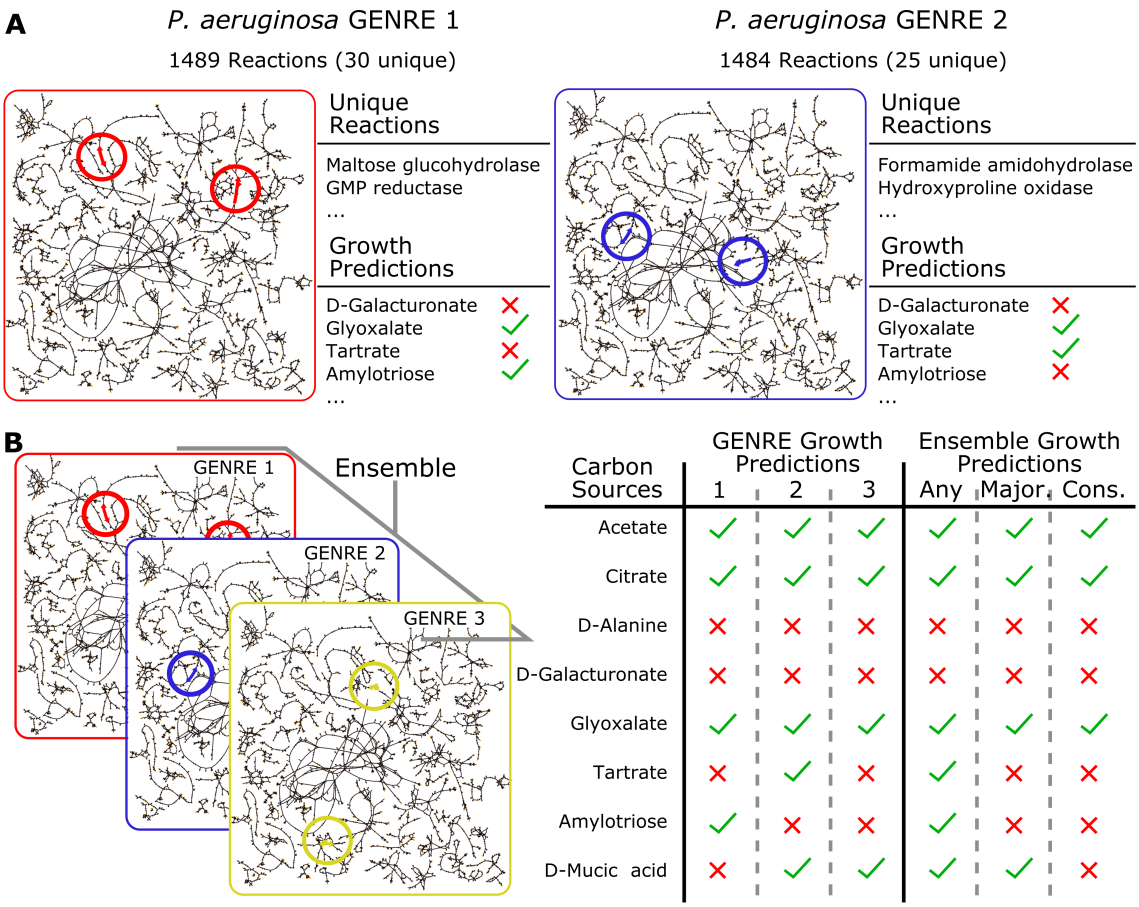


Figure 6.1: **Alternative network structures can be analyzed collectively as an ensemble**. Gap filling a network using the same media conditions, but in different orders, can lead to different network structures. Here we display two networks for *P. aeruginosa* gap filled with two permutations of the same 10 minimal media conditions. We highlight two unique reactions in each, and growth predictions which differ between the two. B. An ensemble can be created by collecting many alternative network structures which are all consistent with available data. Ensemble-level predictions are generated by treating the individual network predictions like votes. We used three qualitatively different decision thresholds: the "any" threshold requires that a single network predict growth; the "majority" threshold requires that a strict majority predict growth; the "consensus" threshold requires all networks within the ensemble to be in agreement. Note that the top five growth conditions result in the same prediction regardless of threshold, while the bottom three conditions result in threshold-dependent outcomes.

tations of these conditions (in this case, there are 120 possible permutations). We gap filled in the order prescribed by the first permutation and then in the order of the second, and compared the resulting networks. We repeated the process 30 times, each time drawing a new set of five random media conditions and gap filling using two random permutations of those five media conditions. We found that even with as few as two media conditions, gap filling in a different order resulted in an average of 25 unique reactions per GENRE (Figure 6.2). As the number of media conditions increased, so did the average difference between the resulting GENRES.

"Global" Gap Filling Provides No Advantages Over a Sequential Approach

We hypothesized that rather than gap filling sequentially, perhaps a "global" gap fill approach would result in more parsimonious, biologically-relevant solutions without the ambiguity associated with changing the gap fill order. We extended our custom gap fill algorithm to identify a minimal set of reactions which could be added to a GENRE to permit growth in multiple media conditions simultaneously (see Materials and Methods). We started with the *P. aeruginosa* UCBPP-PA14 draft GENRE from the Model SEED and repeated the 30 replicates from 2 to 30 media conditions as above, but using the global gap fill approach that we developed (see Materials and Methods; Figure 6.3). We found that this global approach did not identify solutions that were any

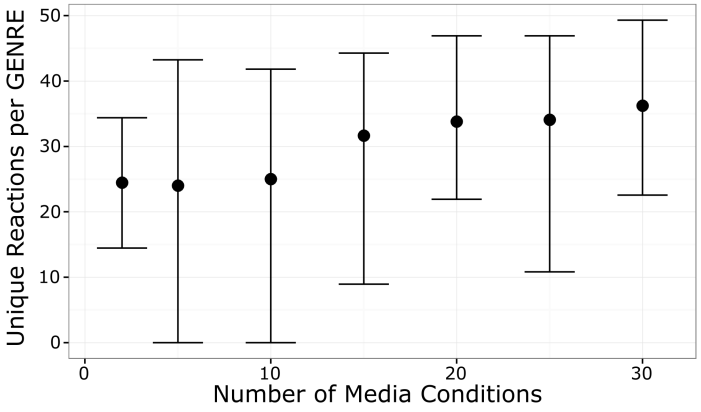


Figure 6.2: **Gap filling in different orders leads to different network structures**. Each error bar indicates an empirical 95% confidence interval from 30 simulations. For a single simulation, a set of media conditions was randomly selected (we simulated sets sizes of 2–30) and we gap filled a GENRE twice, using the same media conditions but in different orders. We compared the resulting pair of GENREs and we found that on average, GENREs within a pair contained an average of 25–35 unique reactions. The average number of unique reactions increased with the number of media conditions used to gap fill.

more parsimonious (Figure 6.3A), and lead to dramatic increases in solve times with increasing media conditions (Figure 6.3B). In order to determine whether the global solution was any more "biologically-relevant", we also compared the ability of the global and sequential approaches to reconstruct a well-curated GENRE for *P. aeruginosa* UCBPP-PA14 called iPAU1129 [Bartell et al. In review]. For each iteration (30 total), we removed 20% of reactions from iPAU1129 and used the sequential and global approaches to gap fill from the universal database using a random selection of five media conditions. The resulting networks were compared to the original iPAU1129, under the assumption that the most biologically-relevant approach would most faithfully reconstructed the curated GENRE, iPAU1129. We found no statistically significant difference between the two approaches (Figure 6.3C) (p-value = 0.63 by two-sided, paired Wilcoxon signed rank test).

Collecting Many Alternative Network Structures into an Ensemble Results in Improved Predictions

Because the sequential gap filling approach produces different results depending on the order of gap filling, we chose to maintain many possible structures resulting from random permutations of the input media conditions rather than select a single GENRE structure for downstream analysis. Not knowing the "true" network structure, we considered each different structure to be a "hypothesis" and analyzed them collectively. For each of 2 to 30 training media conditions we produced 21 GENREs by randomizing the gap fill order (Supplemental Figure 1). We then evaluated each GENRE individually by predicting growth or no growth on 34 test media conditions (17 media conditions which experimentally supported growth and 17 which did not) using flux balance analysis (FBA). We found that each GENRE produced slightly different growth predictions, resulting in some GENREs being more accurate than others (Figure 6.4). In order to generate predictions using the ensemble, we treated each GENREs prediction as a single vote, and pooled the votes using a threshold (Figure 6.1B). We tested three qualitatively different thresholds; "any", "majority", and "consensus". The "any" threshold simply requires that at least one GENRE predict growth in a particular media condition. The "majority" threshold requires greater than half to predict growth, and the "consensus" threshold requires all GENREs to predict growth. We evaluated the growth predictions in terms of accuracy, precision, and recall (see Materials and Methods).

We found that the "majority" threshold led the overall ensemble to achieve average accuracy with respect to the individual GENREs, consistently outperforming the least accurate of the individual GENREs (Figure 4 "Order Only"). The "any" threshold decreased overall accuracy and precision to be worse than any individual GENRE, but increased the recall to match the best individual GENREs. At the other extreme, we found that the "consensus" threshold led to accuracy and precision that matched the very best individual GENREs but diminished recall.

Increasing the Diversity of Network Structures Increases Recall

While different gap fill order does result in different GENRE structures, the differences are relatively small (tens of differences relative to hundreds of reactions overall). In order to span a greater range of potential GENRE structures, we added random weights (drawn from a uniform distribution) to the reactions in the gap filling step (see Materials and Methods). The rationale is that given two pathways of slightly different length but the same biological function, random weights will occasionally favor the longer pathway, thus exploring alternatives that would otherwise be unobserved given a strictly parsimonious procedure. Additionally, each GENRE was reconstructed using a random subset of

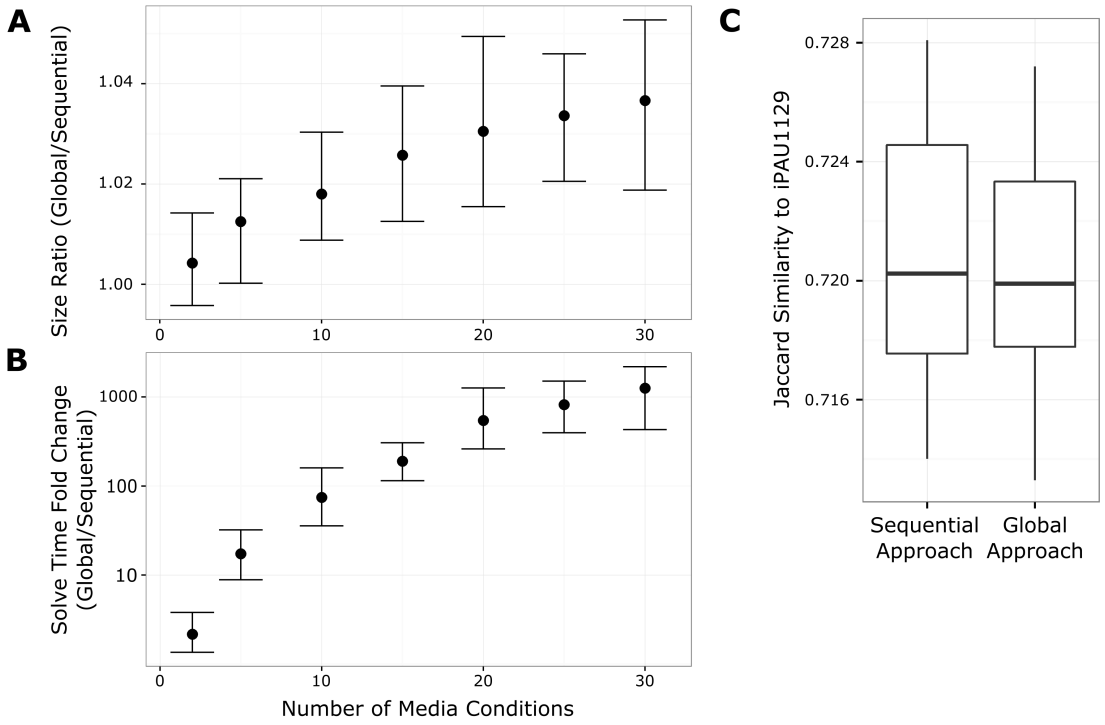


Figure 6.3: **Results of global gap filling approach are no more parsimonious or biologically relevant**. For each number of media conditions, we reconstructed 30 pairs of GENREs. For each pair, a set of media conditions was randomly selected and one GENRE was gap filled sequentially while the other was gap filled using a global approach. A. We found that the GENREs resulting from the global approach were slightly larger than those gap filled using the sequential approach, with an increasing size disparity as the number of media conditions increased. B. The global gap fill approach required significantly more time to run (note the log scale on the y-axis). The solve time increased quadratically with the number of media conditions, such that with 30 media conditions the average solve time for the global approach was ∼1000 times greater than the sequential approach. Error bars in panels A and B represent empirical 95% confidence intervals. C. We compared the ability of the sequential and global approaches to replace reactions removed from a manually-curated GENRE for *P. aeruginosa* UCBPP-PA14, iPAU1129. For each replicate, we removed 20% of the reactions from iPAU1129 and applied the sequential and global gap filling approaches with the same set of randomly selected media conditions. We compared the reaction content of the gap filled GENREs with iPAU1129 using the Jaccard similarity metric. We found that there was no difference between the sequential and global approaches in terms of recovering the removed reactions (p-value = 0.63 by two-sided, paired Wilcoxon signed rank test). Box plots indicate quartiles of the distributions.

only 80% of the reactions from the draft GENRE from the Model SEED. Using this new procedure, we reconstructed ensembles of 21 GENREs using 2 through 30 training media conditions (Supplemental Figure 2). We evaluated the accuracy by predicting growth on the same 34 test media conditions as before. The resulting accuracy, precision and recall of the individual GENREs were essentially the same on average (Figure 4 "Diverse"), but the distribution spanned a much greater range, both positively and negatively. In this case, the "majority" threshold again achieved average behavior with respect to the individual GENREs, outperforming the least accurate individual GENREs (Figure 4 "Diverse"). The "any" threshold tended to achieve the best accuracy and precision, although not quite as good as the best individual GENREs. However, the "any" threshold achieved the best recall, better than the best individual GENREs and better than the recall achieved with a less diverse ensemble.

Accounting for Negative Growth Conditions Greatly Improves Ensemble Accuracy, Precision and Recall

Our experimental growth data for *P. aeruginosa* UCBPP-PA14 included both positive (media conditions which supported growth) and negative results (media conditions which did not support growth). We formulated an optimization-based procedure which allowed us to incorporate information inherent in negative growth conditions into our automated curation (see Materials and Methods). In brief, the optimization problem identifies a minimal number of reactions to "trim" from a GENRE in order to prevent growth on negative media

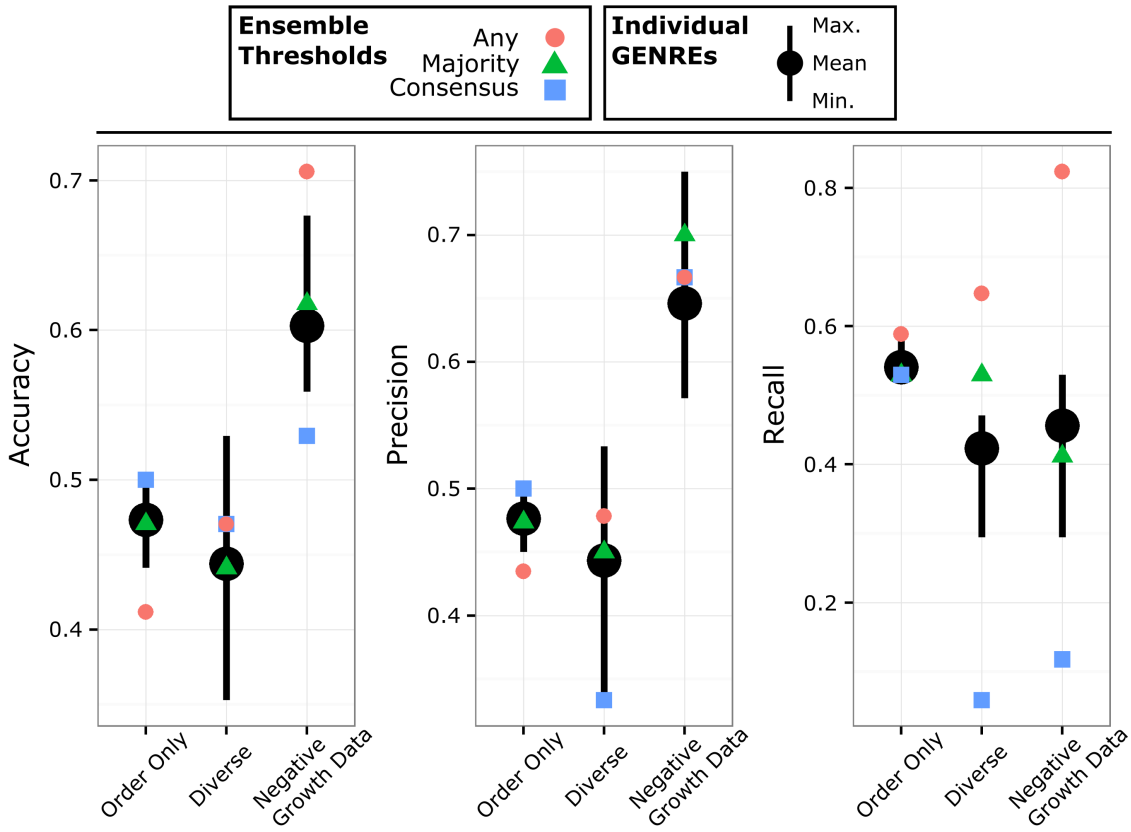


Figure 6.4: **Ensembles generated by gap filling against the same media conditions in different orders**. Using 20 media conditions, we generated 21 GENREs, where each GENRE was gap filled using either: a different order of the same input media conditions ("Order Only"), random weighting of reactions in database and random subsets of reactions from draft ("Diverse"), or a diverse ensemble which also included negative growth data through a trimming step ("Negative Growth Data"). We evaluated the accuracy, precision, and recall of every individual GENRE and of the ensembles by predicting growth on 17 positive media conditions and 17 negative media conditions which were not used during gap filling. The average of the individual GENREs is shown as black points with the maxima and minima as black lines extended above and below. The ensemble predictions using the three different thresholds are shown as red circles "any", green triangles "majority", and blue squares "consensus". Note that there is less ensemble diversity when differences result only from media condition ordering (maxima/minima of "Order Only" compared to "Diverse" or "Negative Growth Data"). Adding additional diversity results in GENREs with both greater and lower accuracy than the best and worst of "Order Only". Addition of the trimming step ("Negative Growth Data") improves overall accuracy and precision by ∼15%. In terms of ensemble thresholds, the "majority" threshold tends to perform similarly to the average of the individual GENREs. The "any" threshold achieves recall as good or better than the best individual GENREs. The "consensus" threshold performs consistently well in terms of accuracy and precision if there is very little diversity in the ensemble ("Order Only").

conditions while maintaining growth on positive media conditions. As before, we generated ensembles of 21 GENREs for 2 through 30 positive media conditions (Supplemental Figure 3). We used random reaction weights, random subsets of 80% of the reactions from the draft GENRE from Model SEED, and this time we selected 10 negative media conditions for each GENRE (distinct from the negative conditions used to assess accuracy) and incorporated them using our trimming procedure. We found that incorporation of the negative media conditions increased the accuracy and precision of both the individual GENREs and the ensembles by ∼15% (Figure 4 "Negative Growth Data"). The "majority" threshold once again approximately tracked the average GENRE accuracy, precision and recall. The "any" threshold achieved accuracy and recall that were often better than the top individual GENREs, with recall exceeding that achieved previously.

Ensembles Achieve Greater Precision or Recall Than Best Individual GENREs When Predicting Essential Genes

We evaluated the ability of ensembles to predict gene essentiality. We generated an ensemble of 51 GENREs, each created by gap filling with a random subset of 25 of the total 47 positive media conditions (53%), 10 randomly-selected negative media conditions from the total of 40 (25%), and 1,210 randomly-selected reactions from a total of 1,512 in the draft GENRE generated by Model SEED (80%). Genes were associated with reactions based on the assigned gene-protein-reaction (GPR) relationships from the draft GENRE. We used an *in silico* representation of CF sputum medium and predicted gene essentiality by removing reactions associated with each gene in turn (according to the GPR logic) and running FBA. We compared the resulting gene essentiality predictions with experimental results [19]. We found that the "majority" threshold resulted in better accuracy and recall than the average of individual GENREs, and drastic improvement over the worst GENREs (Figure 6.5). The "consensus" threshold resulted in a ~20% increase in precision over the best individual GENRE and greater than 100% increase over the worst individual GENRE. Unsurprisingly, the increased precision of the "consensus" threshold comes at the cost of reduced recall. The "any" threshold resulted in lower precision but a ~40% increase in recall over the best individual GENRE and a ~170% increase over the worst individual GENREs.

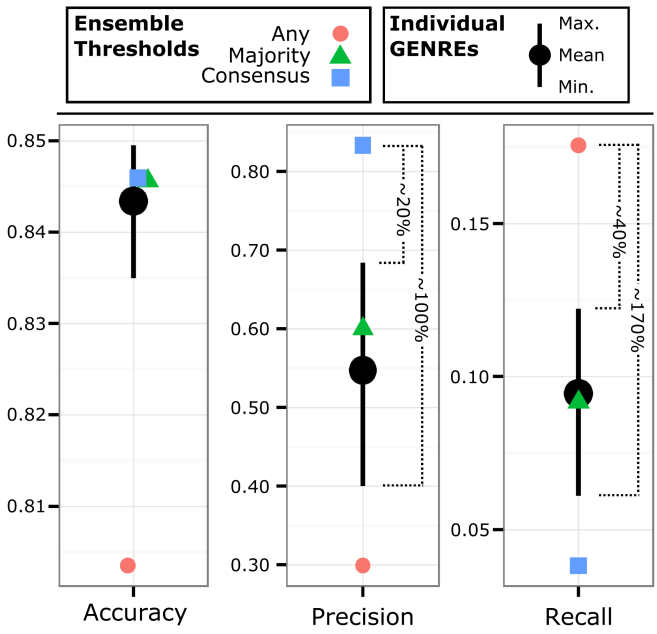


Figure 6.5: **Ensembles outperform individual GENREs when predicting gene essentiality.** We generated an ensemble of 51 GENREs by gap filling against 25 randomly-selected positive growth conditions, 10 negative growth conditions, and 80% of the reactions from the Model SEED draft network. We predicted gene essentiality in CF sputum medium and compared the predictions to *in vitro* gene essentiality data. We found that the "consensus" threshold (blue squares) achieved a ~20% increase in precision over the best individual GENRE and a ~100% increase in precision over the worst individual GENRE. Similarly, the "any" threshold (red circles) achieved a ~40% increase in recall over the best individual GENRE and a ~170% increase over the worst. Note the threshold-dependent tradeoff between precision and recall.

Increasing Ensemble Size Improves Predictions for Small Ensembles

Using the same ensemble of 51 GENREs from above, we examined the effect of ensemble size on predicting essential genes. We sampled with replacement 10,000 small ensembles from among the 51 GENREs for ensemble sizes of 2 through 51. We evaluated the accuracy, precision and recall against the same gene essentiality data set using a "majority" threshold. We found that smaller ensembles were less accurate, less precise, and more variable than larger ensembles (Figures 6.6A and 6.6B). Increasing size improved predictions but with diminishing benefits as the ensemble grew larger. Interestingly, with this "majority" threshold, average recall increased initially, but diminishes again as the ensemble grows larger (Figure 6.6C).

Common Reactions in Ensemble Are Consistent with Manually-Curated Reconstruction

In order to characterize the way gap filling distributes reactions throughout the ensemble, we generated an ensemble of 100 GENREs (Figure 6.7A). Each GENRE was reconstructed using a randomly-selected 80% of the reactions in iPAU1129, and then sequentially gap filled from the independent, universal reaction database using 25 random positive growth conditions. We found that before gap filling, the "correct" reactions from iPAU1129 were initially distributed in a bell curve throughout the ensemble (Figure 6.7B). The vast majority of "correct" reactions were found in 50 or more of the GENREs, and in 80 GENREs on average. In contrast, the "incorrect" reactions (those added by gap filling but which were not in the original iPAU1129) were distributed sporadically,

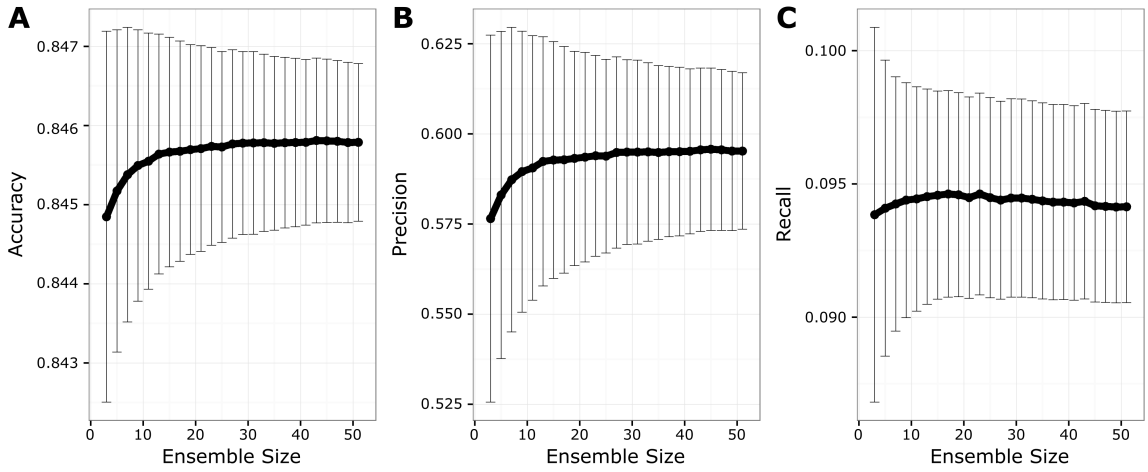


Figure 6.6: **Increasing ensemble size improves performance initially**. Using the same ensemble of 51 GENREs, we used bootstrap sampling to simulate 10,000 ensembles of sizes 2 through 51. We evaluated the performance of each sampled ensemble in terms of accuracy (A), precision (B), and recall (C) on the gene essentiality predictions using the "majority" threshold. We found that accuracy and precision increased sharply until around 15 GENREs, at which point gains were less pronounced with additional GENREs. This result suggests that increasing ensemble size does not infinitely improve ensemble performance.

with the majority being found in 10 or fewer GENREs. After the gap filling step, 65 "correct" reactions were found to have been added to every GENRE, suggesting a core set of "correct" reactions that were required for biomass production in any condition. We observed that the most common reactions (found in 50 or more GENREs) were overwhelmingly "correct" reactions from iPAU1129 (Figure 6.7C). All of these most common reactions (both "correct" and "incorrect") were involved in the production of biomass components, particularly amino acids.

Identifying Small Molecules Which Interact with Unique *Streptococcus* Species

We demonstrate how EnsembleFBA can be implemented in a systems biology workflow. We selected six species from the genus *Streptococcus* which all have growth phenotype data available through a previous study [20]. We reconstructed an ensemble for each species: *Streptococcus mitis*, *Streptococcus gallolyticus*, *Streptococcus oralis*, *Streptococcus equinus*, *Streptococcus pneumoniae* and *Streptococcus vestibularis* (Figure 6.8A). For each species, we generated a draft GENRE using the Model SEED online interface. We generated a diverse ensemble of 21 GENREs from each Model SEED draft, and gap filled each member GENRE using 25 random growth conditions specific to that species. We mapped all genes (translated to protein sequences) from each *Streptococcus* species to small molecule protein binding sequences from DrugBank using NCBI standalone BLASTP and an e-value threshold of 0.001 [21, 22]. For all potential gene targets, we used the ensembles to predict gene essentiality using a "majority" threshold in rich media.

We found 261 small molecules in DrugBank that potentially bind to the products of 169 essential genes (evenly distributed throughout the six species). Many of these small molecules (113) interact with an essential gene in only one of the species, while 44 were predicted to target conserved essential genes in all six species (Figure 6.8B). *S. equinus* was predicted to have the most essential genes interact with unique small molecules while *S. pneumoniae* was not predicted to have essential genes interact with any unique small molecules (Figure 6.8C). As an example of a conserved small molecule interaction, DB04083 (N′-Pyridoxyl-Lysine-5′-Monophosphate) is predicted to interact with essential aspartate aminotransferases in all six species. Alternatively, DB03222 (2′-Deoxyadenosine 5′-Triphosphate) is only predicted to interact with an essential ribonucleotide reductase in *S. gallolyticus*.

To better understand the differences between the metabolic networks which underpin these small molecule screen results, we predicted reaction essentiality in rich media for all six species using a "majority" threshold. We found that several metabolic subsystems were enriched among essential reactions beyond what would be expected from random chance (Figure 6.8D). Some subsystems were enriched in all six species, such as Peptidoglycan biosynthesis, indicating that these reactions related to cell wall biosynthesis are disproportionately essential in all six species. Other subsystems were enriched among essential reactions in a unique species. For example, *S. mitis* is predicted to have a greater proportion of essential reactions related to Amino acid metabolism

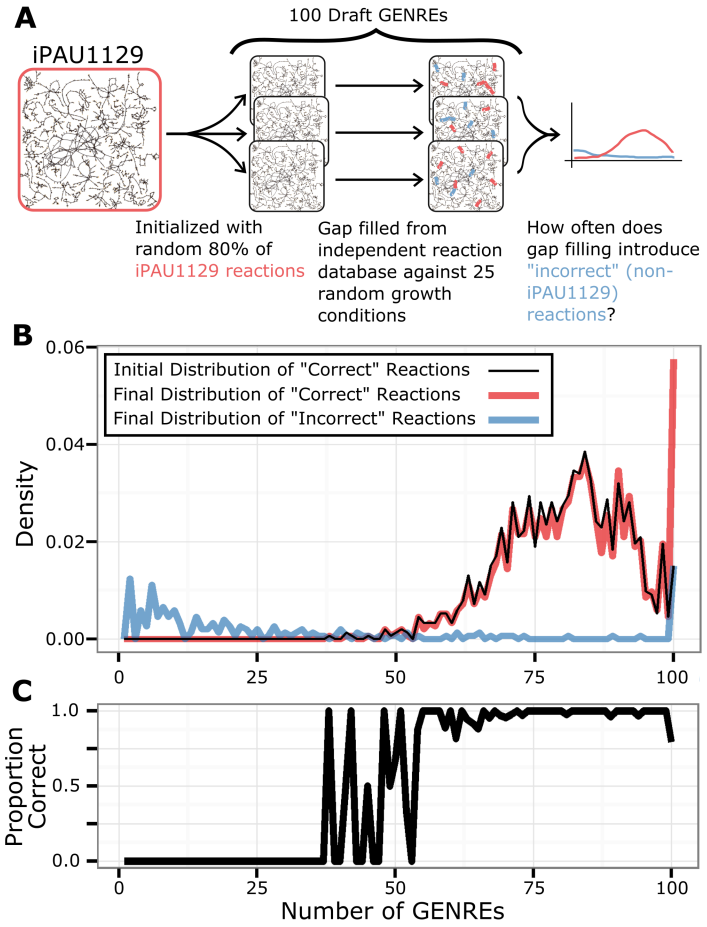


Figure 6.7: **Common reactions in ensemble are consistent with manually-curated reconstruction.** A. We generated an ensemble of 100 GENREs using a randomly-selected 80% of the reactions in iPAU1129, and then sequentially gap filled from the universal reaction database using 25 randomly-select positive growth conditions. B. The distribution of reactions throughout the ensemble is displayed as the proportion of reactions (y-axis) which are found in a given number of GENREs (x-axis). "Correct" reactions are whose which are found in the manually-curated iPAU1129, while "incorrect" reactions are those which are added during gap filling but not found in iPAU1129. We observed that there is a common set of reactions which were found in all 100 GENREs. The majority of this common set are "correct" (88 reactions) while 23 are "incorrect". C. The common reactions (found in 50 or more GENREs) consist of a greater proportion of "correct" reactions. "Incorrect" reactions tend to be uncommon.

than other species, perhaps indicating that *S. mitis* has less redundancy in those pathways than the other six species. Essential reactions related to Butanoate metabolism were most enriched in *S. pneumoniae*, while essential reactions in Lysine degradation were most enriched in *S. equinus*. Interestingly, reactions associated with core metabolic functions (e.g. Amino acid biosynthesis, Valine and leucine biosynthesis, Phenylalanine biosynthesis) were not equally enriched among essential reactions for all species.

6.5 Discussion

Genome-scale metabolic network reconstructions (GENREs) have been used for decades to assemble information about an organism's metabolism, to formally analyze that information, and in so doing, to make predictions about that organism's behavior in unobserved or unobservable contexts. A major barrier preventing more widespread use of GENREs, particularly in non-model organisms, is the extensive time and effort required to produce a high-quality GENRE. Many automated approaches have been developed which reduce this time requirement (e.g. Model SEED, GLOBUS, CoReCo, RAVEN) [13–16]. We demonstrate that gap filling—although our results apply to many automated curation approaches—can lead to many potential GENRE structures depending on the ordering of the input data. Rather than arbitrarily selecting a single GENRE from among many possible networks (which are all reasonably consistent with the available data), we found that collecting many GENREs into an ensemble improved the predictions that could be made. We call this approach "EnsembleFBA" and emphasize that ensembles are a useful tool for dealing with uncertainty in network structure. We demonstrated how ensemble diversity impacts predictions. We show that EnsembleFBA correctly identifies many more essential genes in the model organism *P. aeruginosa* UCBPP-PA14 than the best individual GENREs. We showcase how EnsembleFBA can be utilized in a systems biology workflow by predicting how small molecules interact with different essential genes in six Streptococcus species. Ensembles increase the quality of predictions without incurring months of manual curation effort, thus making GENRE approaches more practical for applications which require predictions for many non-model organisms. We have provided code to facilitate the creation and analysis of ensembles of GENREs.

Gap filling is a common step during the GENRE curation process, both for manually- and automatically-curated GENREs [5]. We used a linear (rather than binary) gap filling algorithm to expand GENREs so that they are capable of producing biomass *in silico* on growth media which supports growth of the organism *in vitro*. Gap filling algorithms suggest parsimonious reaction sets from some "universal" biochemical database which, if added to a GENRE, will allow growth in the

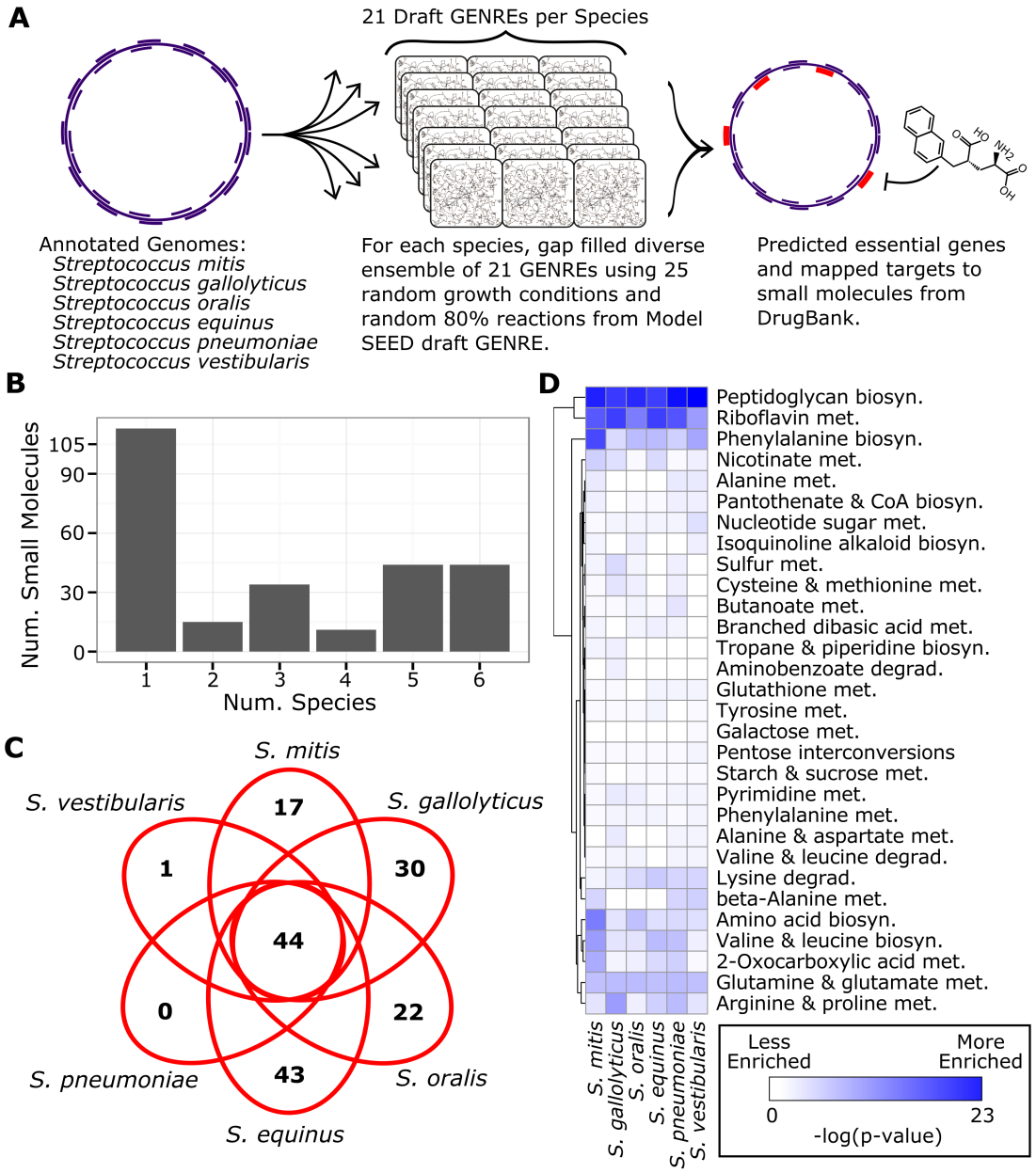


Figure 6.8: **EnsembleFBA predicts unique essential genes targets of small molecules in six Streptoccocus species**. A. We reconstructed ensembles of 21 GENREs for six *Streptococcus* species based on draft GENREs from the Model SEED and 25 growth conditions. From among all genes within all six species, we identified with potential binding interactions with small molecules from the DrugBank database, and used the ensembles to predict the essentiality of those genes. We found 261 small molecules with potential binding to essential gene products. B. Many small molecules interact with an essential gene in only one species, while a core set of 44 small molecules interact with essential genes in all six species. C. Distribution of small molecule interactions with essential genes, unique and conserved among the six species. Note that 44 small molecules interact with essential genes in all six species. *S. equinus* is predicted to have essential genes uniquely interact with 43 small molecules, while *S. pneumoniae* is predicted to not have any essential genes which interact with unique small molecules. D. Subsystem enrichment among essential reactions by species. We predicted reaction essentiality for all six species in rich media and then calculated a p-value indicating the likelihood of observing each subsystem among the essential reactions given the total number of reactions associated with that subsystem. For clarity we display the $-log$(p-value), where darker colors indicate greater enrichment (i.e. a disproportionate number of reactions in that subsystem are predicted to be essential). Note that some subsystems are enriched among essential reactions in all six species (e.g. Peptidoglycan biosynthesis) while others are uniquely enriched in a specific species (e.g. Phenylalanine biosynthesis in *S. mitis*).

new environment [6, 17, 18, 23]. Often, multiple reaction sets can enable growth, so some heuristics are needed to select a final solution. Sometimes gene homology metrics are used to select a solution, such that genes which catalyze the suggested reactions are compared to the current genome, and the reaction set with the best matches in the current genome are selected as the final solution. When validated, these solutions can lead to re-annotation of the genome [6]. During automated curation, there is less opportunity for extensive validation, and so the first or the most parsimonious solution is selected. As we demonstrated, the order of gap filling can change the final outcome, thus producing GENREs with different structures from the exact same input data (Figure 6.2). Under these circumstances, it is difficult to know which solution is most correct without additional data.

A possible way around this issue of gap fill order is to remove the sequential nature of gap filling entirely and use a global gap filling approach. We demonstrate that not only is such a global approach much slower (quadratic increases in solution time as growth media conditions are added), but the solutions are no more parsimonious or biologically relevant (Figure 6.3). Alternatively, we found that two additional innovations improved the predictions that could be obtained from automatically-generated GENREs: the addition of negative growth conditions and the collection of multiple GENREs into ensembles.

Negative growth conditions have not been extensively incorporated into GENRE curation. To our knowledge, only one group has developed an approach for removing reactions in order to prevent growth in specific conditions [23]. In that case, the reactions were not removed from the GENRE, but rather, prevented from carrying flux under particular conditions. This approach was supported by a biological justification that certain enzymes may not be functional under certain conditions [23]. Our approach is different in that it seeks to produce a single GENRE structure that is consistent with all available data, positive and negative. We achieved this by removing a minimal reaction set to simultaneously prevent growth in negative conditions and allow growth in positive growth conditions (see Materials and Methods). By utilizing this untapped source of information, we found that average GENRE accuracy increased by ∼15% (Figure 6.4). Automatically incorporating negative growth conditions is a little-explored area that has the potential to make better use of growth screening data.

Ensembles have been used for many years in the machine learning community to leverage the strengths of many different models to improve predictions [24, 25]. Ensembles have been used previously to analyze GENREs from a kinetic standpoint [26]. Because kinetic parameters are usually unknown for an entire genome-scale network, ensembles of kinetic parameters are generated such that all parameter sets lead to the same steady state [26]. In this way, ensembles can represent the space of allowable kinetic parameters. Our approach to generating ensembles is different in that we attempt to represent the space of allowable GENRE structures rather than kinetic parameters.

Ensembles provide a significant advantage over individual GENREs by tuning for specific results with defined decision thresholds (Figures 6.4 and 6.5). Consistently, by using the "any" threshold, recall can be made to equal or exceed the best individual GENREs. This result makes sense, considering that different network structures will result in different growth or gene essentiality predictions. By accepting any essential gene prediction from among the constituent GENREs, we cast a wider net and capture many more of the true essential genes and growth conditions. The fact that many individual GENREs contribute unique but true predictions suggests that each GENRE recapitulates elements of the "true" network structure (Figure 6.7). Similarly, by using the "majority" threshold, the ensemble predictions perform like the average GENRE (Figures 6.4 and 6.5). By requiring a majority of GENREs to agree, the ensemble guards against poor predictions and, in most cases, outperforms the worst individual GENREs. Finally, if precision is the overall goal, a "consensus" threshold provides confidence that the majority of positive predictions are true positives (Figure 6.5).

We observed that ensemble performance is limited by the quality of the GENREs which form the ensemble. The choice of decision threshold ("any", "majority", or "consensus") did not consistently improve overall accuracy of the ensemble. However, by improving the individual GENREs using negative information, the overall ensemble accuracy improved dramatically (Figure 6.4). Also, it should be noted that the computational burden required by ensembles will always be greater than the burden of a single GENRE. For all the examples in this study, computational burden scales linearly with the number of GENREs in the ensemble (ensemble of size N GENREs will require N times longer to calculate FBA solutions) which is a modest expectation in practice. Other applications, like predicting species interactions, would not scale linearly if all possible pairs of GENREs between two ensembles were simulated.

Increasing ensemble diversity impacted ensemble recall, but did not have an obvious effect on overall accuracy. Some degree of diversity is required in order to gain any advantage through an ensemble representation. In the "Order Only" ensemble (generated simply by chang-

ing the order of gap filling; Figure 6.4) there were only small differences between any of the GENREs so it was difficult to improve on the best GENRE. By injecting greater diversity through random weights and random subsets of the data, we observed much greater variation in individual GENRE performance (both positively and negatively), but the average accuracy was the same as the low diversity ensemble (Figure 6.4). The advantage of diversity is in casting a wide net and thus improving ensemble recall, particularly when combined with an "any" decision threshold. In practice, the choice to increase diversity or not will depend on the goals of the analysis. If the goal is to generate many candidate essential genes or media conditions, then more diversity will be advantageous. If the goal is to generate fewer, more confident predictions, then minimizing diversity will be most effective.

EnsembleFBA is easily integrated into systems biology workflows. As an example, a current challenge in systems biology is to identify species-specific drug targets so that therapies will not disrupt the healthy microbiome structure [27, 28]. We reconstructed ensembles for six *Streptococcus* species by gap filling with growth phenotype data, we predicted essential genes and mapped those genes to potential small molecule binding partners within a matter of hours, and can have more confidence in the quality of the gene essentiality predictions than if we were to work with single GENREs for each species (Figure 6.8). The process scales well with the number of species, such that 12 or 100 species would not take significantly longer than six, and the quality of the predictions is maintained with scale. It is interesting to note that among *Streptococcus* species, there are generally small molecules which can be selected to uniquely interact with essential genes in a single species, and other small molecules which interact with conserved essential genes (Figure 6.8C). The observed interactions between essential genes and small molecule ligands are species-specific because of differences in network structure which lead to some metabolic subsystems being disproportionately represented among essential reactions (Figure 6.8D). In the search for species-specific drug targets, it is important to consider, not only the presence or absence of a particular gene, but also the role of that gene in the broader network context, and improved systems biology tools such as EnsembleFBA can help to elucidate that context with greater confidence.

Gap filling is not the only GENRE reconstruction approach that produces many possible solutions. Likelihood-based gap filling produces a distribution of possible annotations for each gene in a genome, assigning a probability to each [14, 16, 29]. Network structure is then based on maximizing the likelihood over all possible solutions. Ensembles could be generated easily using this type of framework by sampling many alternative solutions around the maximum likelihood. Indeed, it may be beneficial to create an ensemble using GENREs reconstructed using several different methods. We suggest that there are many possible ways to generate ensembles such that they will allow researchers to generate better predictions about under-studied organisms.

Finally, we foresee ensembles playing an important role beyond improving predictions, for example, in experimental design and model reconciliation. Within a diverse ensemble, many possible network structures are represented, and it is expected that some structures will be closer to the truth than others. We suggest that ensembles can be leveraged to design an optimal series of experiments to weed out the most incorrect network structures. For instance, such an approach could select the most differentiating carbon sources to experimentally test, or the most differentiating essential genes. Ensemble-guided experimental design could save time and experimental resources. Model reconciliation is another field that could benefit from ensembles [8, 30]. Given GENREs for two different species, reconciliation is the process of removing systematic differences from the two GENREs so that any differences which remain are due to biology alone. Systematic differences often result from arbitrary choices during the process of reconstruction. Ensembles could be used to automate the reconciliation process by representing the space of possible GENREs for each species and the reconciled versions would be the two models from the two spaces that are most similar to each other. Thus, ensembles have potential to improve other tasks than prediction, including experimental design and mapping the space of GENRE structures for tasks like reconciliation.

6.6 Materials and Methods

Code and Data Availability

All data, Matlab (Natick, MA, USA) implementations of algorithms, Matlab simulation scripts, results files and figure generation scripts are publicly available in our online repository: github.com/mbi2gs/ensembleFBA

Data Sources

All biochemical reference data was obtained from the Model SEED database (github.com/ModelSEED/ModelSEEDDatabase).
The metabolic reaction and compound databases were parsed and formatted for use in Matlab using a custom Python script available in our reposi-

tory ("format_SEED_data.py"). A draft network for *P. aeruginosa* UCBPP-PA14 was automatically generated using the Model SEED web service (modelseed.org/genomes/). Similarly, draft networks were generated for *Streptococcus mitis* ATCC 6249, *Streptococcus gallolyticus* ICDDRB-NRC-S3, *Streptococcus oralis* ATCC 49296, *Streptococcus equinus* AG46, *Streptococcus pneumoniae* (PATRIC ID 1313.5731), and *Streptococcus vestibularis* 22-06 S6.

Representations of media conditions (including minimal media and cystic fibrosis sputum medium), and biomass representations were drawn from previous GENRE analyses of *Pseudomonas aeruginosa* [31, 32].

P. aeruginosa PA14 essential genes in cystic fibrosis sputum medium were experimentally identified previously [19].

A manually curated, and thoroughly validated GENRE of *P. aeruginosa* UCBPP-PA14 called iPAU1129 was developed previously [Bartell et al. In review], along with Biolog growth screen data for *P. aeruginosa* UCBPP-PA14 indicating many media conditions in which this strain will and will not grow.

Growth phenotype data for six *Streptococcus* species was obtained from the file "Supplementary Data 1" of [20]. Small molecule amino acid binding target sequences were downloaded from the DrugBank website (http://www.drugbank.ca/) [21]. After identifying homologous genes to the target sequences using BLASTP [22], we used a custom python script to parse the results for input into Matlab ("listPossibleTargets.py", available in repository).

Linear Gap Filling

We implemented a linear (as opposed to binary) gap filling algorithm in Matlab, based on the algorithms FASTGAPFILL and FastGapFilling [17, 18]. We used the Gurobi solver version 6.0.5 for all optimization tasks (Gurobi, Houston, TX, USA). To begin, we provide the algorithm with a universal database of metabolic reactions U, a universal database of exchange reactions X, a biomass reaction, and a set of growth conditions formatted as lower bounds on exchange reactions. The algorithm identifies a set of reactions from U and X that allow flux through the biomass reaction under all growth conditions. The algorithm is implemented as a linear program (LP) that minimizes the sum of the absolute value of all fluxes through U and X. The optimization problem takes the form:

$$
\begin{array}{lll}
\min_z & \sum r_u z_u + \sum r_x z_x & \\
\text{s.t.} & Uv + Xw = 0 & (1) \\
& lb_{u,i} \le v_i \le ub_{u,i} \quad \forall i \in [1, N_u] & (2) \\
& lb_{x,i} \le w_i \le ub_{x,i} \quad \forall i \in [1, N_x] & \dots \\
& -z_{u,i} \le v_i \le z_{u,i} \quad \forall i \in [1, N_u] & (3) \\
& -z_{x,i} \le w_i \le z_{x,i} \quad \forall i \in [1, N_x] & \dots \\
& z_i \ge 0 \quad \forall i \in [1, N_u + N_x] & (4) \\
& v_{biomass,gc=j} \ge 0.05 \quad \forall j \in [1, N_{gc}] & (5) \\
& z_{u,i} \ge C z_{u,i} \quad \forall i \in [1, N_u] & (6) \\
& z_{x,i} \ge C z_{x,i} \quad \forall i \in [1, N_x] & \dots
\end{array}
$$

Where: U is the universal reaction library (as a stoichiometric matrix); X is the universal exchange library (same metabolites as U); N_u and N_x are the number of reactions in U and X, respectively; N_{gc} is the number of growth conditions; v is the vector of fluxes through U; w is the vector of fluxes through X; $lb_{u,i}$, $ub_{u,i}$, $lb_{x,i}$, and $ub_{x,i}$ are the lower and upper bounds on v_i and w_i, respectively; $v_{biomass,gc=j}$ is the flux through the biomass reaction under growth condition j; $Cz_{u,i}$ and $Cz_{x,i}$ are variables that can force reactions from U and X to be included; z is a continuous variable which acts as a constraint on the absolute values of elements in v and w; r is an optional weight on z which can be randomized.

In order to incorporate genome annotations from a specific organism, we force the inclusion of all associated reactions from those annotations using the Cz_u variables. Note that unlike a binary optimization, the LP minimizing the sum of the absolute flux values through U and X does not necessarily result in a solution with the fewest reactions, but rather the solution which requires the minimum sum of the absolute values of the fluxes through it. The LP here can be extended to utilize multiple growth conditions simultaneously (global approach) by duplicating the U and X matrices, once for each growth condition, but minimizing a single set of z variables across all conditions. To gap fill using multiple growth conditions sequentially, we gap fill using the first growth condition, incorporate the solution into the GENRE, then repeat the process for all growth conditions. Our Matlab function "expand()" implements this optimization problem.

Incorporating Negative Growth Conditions by Trimming Reactions

We implemented a binary optimization problem to trim minimal reactions from a GENRE in order to prevent growth under negative growth conditions while simultaneously maintaining growth in the positive growth conditions. As input to the algorithm, we provide a GENRE, and a set of both positive and negative growth

conditions. We chose to run FBA first to identify mismatches between the computational predictions and the *in silico* data (negative growth conditions erroneously predicted to support growth *in silico*). Having identified those, we then ran FBA on all the positive growth conditions to identify the top five with flux distributions most similar to the flux distribution of the negative growth condition. The GENRE and the selected growth conditions are passed to the trimming problem, which takes the form:

$$
\begin{array}{lll}
\max_y & \sum r_u y_u + \sum r_x y_x & \\
\text{s.t.} & Uv + Xw = 0 & (1) \\
& y_{u,i} lb_{u,i} \le v_i \le y_{u,i} ub_{u,i} \quad \forall i \in [1, N_u] & (2) \\
& y_{x,i} lb_{x,i} \le w_i \le y_{x,i} ub_{x,i} \quad \forall i \in [1, N_x] & \dots \\
& v_{biomass,gc=j} \ge 0.05 \quad \forall j \in [1, N_{gc}] & (3) \\
& max(v_{biomass,ngc=j}) = 0 \quad \forall j \in [1, N_{ngc}] & (4) \\
& y_u, y_x \in [0, 1] &
\end{array}
$$

Where: U is the set of reactions from the GENRE (as a stoichiometric matrix); X is the set of exchange reactions from the GENRE (same metabolites as U); N_u and N_x are the number of reactions in U and X, respectively; v is the vector of fluxes through U; w is the vector of fluxes through X; $lb_{u,i}$, $ub_{u,i}$, $lb_{x,i}$, and $ub_{x,i}$ are the lower and upper bounds on v_i and w_i, respectively; $v_{biomass,gc=j}$ is the flux through the biomass reaction under growth condition j; y_u and y_x are arrays of binary variables which determine inclusion of reactions from U and X in the network; r is an optional weight on y which can be randomized.

The term $max(v_{biomass,ngc=j}) = 0$ requires that the maximum possible flux through the biomass reaction for the negative growth conditions is constrained to be zero. In order to implement this constraint, we took advantage of duality theory as has been done previously [7]. Specifically, the optimal objective value of the dual of a linear program will equal the optimal value of the primal. By constraining the primal and dual objectives to equal each other, we can ensure that the flux through the biomass objective is maximized. We can replace the term $max(v_{biomass,ngc=j}) = 0$ with the following constraints:

$$
\begin{array}{ll}
v_{biomass,ngc=j} = \lambda_{ub} uby - \lambda_{lb} lby & (4.1a) \\
S^T \lambda_{mets} + \lambda_{ub} - \lambda_{lb} = c & (4.2) \\
\lambda_{ub}, \lambda_{lb} \ge 0 & (4.3)
\end{array}
$$

Where λ_{mets}, λ_{ub} and λ_{lb} are the dual vectors associated with the metabolites, upper and lower bounds of the primal problem. c is a binary vector indicating the objective reaction in U. Note that the terms $\lambda_{ub} uby$ and $\lambda_{lb} lby$ are quadratic, requiring a multiplication of the binary inclusion variable y with the dual variables. Because y is a binary variable, in this case the quadratic constraints can be converted to linear constraints through the substitution of additional variables:

$$
\begin{array}{ll}
t_{ub} \le L_\lambda y_i & (4.4) \\
t_{ub} \ge 0 & (4.5) \\
t_{ub} \le \lambda_{lb,i} ub_i & (4.6) \\
t_{ub} \ge \lambda_{lb,i} ub_i - L_\lambda (1 - y_i) & (4.7)
\end{array}
$$

Where t_{ub} is a stand-in for the product $\lambda_{ub} uby$ and L is a large number greater than or equal to the upper bound on $\lambda_{ub} uby$ (e.g. 1000). Similar constraints would be produced for the product $\lambda_{lb} lby$. The quadratic constraint above (constraint 4.1a) can then be replaced by a linear constraint:

$$
v_{biomass,ngc=j} = t_{ub} ub - t_{lb} lb \qquad (4.1b)
$$

Our Matlab function "trim_active()" implements this optimization problem.

Iterative Approach to Reconstructing GENREs Consistent with All Growth Screening Data

We implemented an iterative algorithm to integrate the LP expansion step with the binary trimming step. The algorithm first applies the expansion step to produce a GENRE that is capable of growing in all positive growth conditions. Next, the algorithm checks for negative growth conditions that allow for biomass flux and for any that do, applies the trim step as described above. The algorithm iterates between the expand and trim steps until either a completely consistent GENRE structure is identified, or it reaches a maximum attempts limit. A single attempt is completed if the GENRE structure is not yet consistent with the input growth conditions but stops making progress (stuck in a local optimum). In this case, a random reaction is removed from the GENRE and the search is re-initiated. If the maximum attempts limit is reached, the algorithm removes any negative growth conditions that are inconsistent with the positive growth conditions, and returns the final GENRE. This iterative algorithm is implemented in our Matlab function "build_network()".

Predicting Growth and Essential Genes

Growth media were simulated by setting the lower bounds on exchange reactions for the appropriate nutrients to negative values. The uptake of carbon source(s) limited the final flux through biomass. "Growth" was

determined by maximizing flux through the biomass objective. We predicted growth if a positive, non-zero flux could be achieved through biomass. Gene knockouts were simulated by generating a new GENRE which was missing the reactions dependent on the knocked-out gene. The reaction-gene dependence was determined by evaluating the binary logic of the GPRs provided by Model SEED. Our custom script to evaluate GPR logic is "simulateGeneDeletion()".

We evaluated the growth predictions in terms of accuracy (TP + TN) / (TP + FP + TN + FN), precision (TP / TP + FP), and recall (TP / TP + FN) where TP = number of true positives, FP = the number of false positives, TN = the number of true negatives and FN = the number of false negatives. Precision indicates the fraction of positive predictions which are true positives. Recall indicates the fraction of positive events which were correctly predicted by the method.

Predicting Small Molecule Interactions

We downloaded the Drug Target Sequences for small molecules in FASTA format from DrugBank [21]. Using NCBI standalone BLASTP and an e-value cutoff of 0.001, we identified homologous sequences in all six Streptococcus proteomes [22].

Metabolic Subsystem Enrichment

We downloaded KEGG subsystem annotations for the reactions in the Model SEED database ("KEGG.pathways.tsv"). After predicting essential reactions for each *Streptococcus* species, we used the hypergeomtric distribution to calculate the probability of drawing k essential reactions and finding that x or more are annotated with subsystem j, from a population of size M reactions, of which N are annotated with subsystem j.

Computational Resources

The majority of our reconstructions and simulations were performed on a 64-bit Dell Precision T3600 Desktop computer with 32 GB RAM and eight 3.6 GHz Intel Xeon CPUs, running Windows 7. Incorporating negative growth information often lead to longer reconstruction times (sometimes 2 hours per GENRE) due to the binary optimization step. To accelerate the reconstruction time while incorporating negative growth information, we used the University of Virginia High Performance Computing Cluster.

Generating Ensembles and Making Predictions Using Our Software

Our Matlab scripts for generating an ensemble (using the gap filling approach described in this work) and for analyzing an ensemble are freely available in a github repository (see Code and Data Availability). The Gurobi solver is required, in addition to our Matlab scripts. We have also included a tutorial script to guide the user through the necessary steps to generate and analyze an ensemble ("test_eFBA.mat").

6.7 Acknowledgments

The authors would like to thank Jennifer Bartell and Anna Blazier for providing a copy of iPAU1129 and the accompanying carbon source utilization data.

6.8 References

[1] Caspi R et al. "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases". In: *Nucleic Acids Research* 38.Database (Jan. 2010), pp. D473–D479. DOI: 10.1093/nar/gkp875.

[2] Jeong H et al. "The large-scale organization of metabolic networks." In: *Nature* 407.6804 (Oct. 2000), pp. 651–4. DOI: 10.1038/35036627.

[3] Oberhardt MA, Palsson BØ, and Papin JA. "Applications of genome-scale metabolic reconstructions." In: *Molecular systems biology* 5 (Jan. 2009), p. 320. DOI: 10.1038/msb.2009.77.

[4] McCloskey D, Palsson BØ, and Feist AM. "Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli." In: *Molecular systems biology* 9.661 (Apr. 2013), p. 661. DOI: 10.1038/msb.2013.18.

[5] Thiele I and Palsson BØ. "A protocol for generating a high-quality genome-scale metabolic reconstruction." In: *Nature protocols* 5.1 (2010), pp. 93–121. DOI: 10.1038/nprot.2009.203.

[6] Reed JL et al. "Systems approach to refining genome annotation." In: *Proceedings of the National Academy of Sciences of the United States of America* 103.46 (Nov. 2006), pp. 17480–4. DOI: 10.1073/pnas.0603364103.

[7] Burgard AP, Pharkya P, and Maranas CD. "Optknock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization". In: *Biotechnology and Bioengineering* 84.6 (2003), pp. 647–657. DOI: 10.1002/bit.10803.

[8] Bartell JA et al. "Comparative Metabolic Systems Analysis of Pathogenic Burkholderia". In: *Journal of Bacteriology* 196.2 (2014), pp. 210–226. DOI: 10.1128/JB.00997-13.

[9] Stolyar S et al. "Metabolic modeling of a mutualistic microbial community." In: *Molecular systems biology* 3.92 (2007), p. 92. DOI: 10.1038/msb4100131.

[10] Zomorrodi AR and Maranas CD. "OptCom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities." In: *PLoS computational biology* 8.2 (Feb. 2012), e1002363. DOI: 10.1371/journal.pcbi.1002363.

[11] Human Microbiome Project Consortium T. "Structure, function and diversity of the healthy human microbiome". In: *Nature* 486.7402 (2012), pp. 207–214. DOI: 10.1038/nature11234.

[12] Lundberg DS et al. "Defining the core Arabidopsis thaliana root microbiome". In: *Nature* 488.7409 (Aug. 2012), pp. 86–90. DOI: 10.1038/nature11237.

[13] Henry CS et al. "High-throughput generation, optimization and analysis of genome-scale metabolic models." In: *Nature biotechnology* 28.9 (Sept. 2010), pp. 977–82. DOI: 10.1038/nbt.1672.

[14] Plata G et al. "Global probabilistic annotation of metabolic networks enables enzyme discovery." In: *Nature chemical biology* 8.october (Sept. 2012). DOI: 10.1038/nchembio.1063.

[15] Agren R et al. "The RAVEN Toolbox and Its Use for Generating a Genome-scale Metabolic Model for Penicillium chrysogenum". In: *PLoS Computational Biology* 9.3 (2013). DOI: 10.1371/journal.pcbi.1002980.

[16] Pitkänen E et al. "Comparative genome-scale reconstruction of gapless metabolic networks for present and ancestral species." In: *PLoS computational biology* 10.2 (2014), e1003465. DOI: 10.1371/journal.pcbi.1003465.

[17] Thiele I, Vlassis N, and Fleming RMT. "FASTGAPFILL: efficient gap filling in metabolic networks". In: *Bioinformatics* 30.17 (Sept. 2014), pp. 2529–2531. DOI: 10.1093/bioinformatics/btu321.

[18] Latendresse M. "Efficiently gap-filling reaction networks". In: *BMC Bioinformatics* 15.1 (2014), p. 225. DOI: 10.1186/1471-2105-15-225.

[19] Turner KH et al. "Essential genome of Pseudomonas aeruginosa in cystic fibrosis sputum". In: *Proceedings of the National Academy of Sciences* 112.13 (Mar. 2015), pp. 4110–4115. DOI: 10.1073/pnas.1419677112.

[20] Plata G, Henry CS, and Vitkup D. "Long-term phenotypic evolution of bacteria". In: *Nature* (2014). DOI: 10.1038/nature13827.

[21] Wishart DS. "DrugBank: a comprehensive resource for in silico drug discovery and exploration". In: *Nucleic Acids Research* 34.90001 (Jan. 2006), pp. D668–D672. DOI: 10.1093/nar/gkj067.

[22] Camacho C et al. "BLAST+: architecture and applications". In: *BMC Bioinformatics* 10.1 (2009), p. 421. DOI: 10.1186/1471-2105-10-421.

[23] Kumar VS and Maranas CD. "GrowMatch: An automated method for reconciling in silico/in vivo growth predictions". In: *PLoS Computational Biology* 5 (2009). DOI: 10.1371/journal.pcbi.1000308.

[24] Opitz D and Maclin R. "Popular Ensemble Methods: An Empirical Study". In: *Journal of Artificial Intelligence Research* 11 (1999), pp. 169–198.

[25] Breiman L. "Random Forests". In: *Machine Learning* 45.1 (2001), pp. 5–32. DOI: 10.1023/A:1010933404324.

[26] Tran LM, Rizk ML, and Liao JC. "Ensemble Modeling of Metabolic Networks". In: *Biophysical Journal* 95.12 (Dec. 2008), pp. 5606–5617. DOI: 10.1529/biophysj.108.135442.

[27] Round JL and Mazmanian SK. "The gut microbiota shapes intestinal immune responses during health and disease". In: *Nature Reviews Immunology* 9.5 (May 2009), pp. 313–323. DOI: 10.1038/nri2515.

[28] McCaughey LC et al. "Efficacy of species-specific protein antibiotics in a murine model of acute Pseudomonas aeruginosa lung infection". In: *Scientific Reports* 6 (July 2016), p. 30201. DOI: 10.1038/srep30201.

[29] Benedict MN et al. "Likelihood-Based Gene Annotations for Gap Filling and Quality Assessment in Genome-Scale Metabolic Models". In: *PLoS Computational Biology* 10 (2014), e1003882. DOI: 10.1371/journal.pcbi.1003882.

[30] Oberhardt Ma, Puchałka J, Santos VaPM dos, and Papin Ja. "Reconciliation of genome-scale metabolic reconstructions for comparative systems analysis". In: *PLoS Computational Biology* 7.3 (2011). DOI: 10.1371/journal.pcbi.1001116.

[31] Oberhardt MA et al. "Genome-scale metabolic network analysis of the opportunistic pathogen Pseudomonas aeruginosa PAO1." In: *Journal of bacteriology* 190.8 (Apr. 2008), pp. 2790–803. DOI: 10.1128/JB.01583-07.

[32] Oberhardt MA, Goldberg JB, Hogardt M, and Papin JA. "Metabolic Network Analysis of Pseudomonas aeruginosa during Chronic Cystic Fibrosis Lung Infection". In: *Journal of Bacteriology* 192.20 (Oct. 2010), pp. 5534–5548. DOI: 10.1128/JB.00900-10.

Chapter 7

Systems-level metabolism of the altered Schaedler flora, a complete gut microbiota

The text for this chapter has been accepted as a research article here:

Biggs MB, Medlock GL, Moutinho Jr. TJ, Lees HJ, Swann JR, Kolling GL$^{\delta}$, Papin JA$^{\delta}$. (2016). Systems-level metabolism of the altered Schaedler flora, a complete gut microbiota. *ISME Journal*, in press.

δ Corresponding authors.

7.1 Context

Around the end of my first year as a graduate student, Dr. Papin, knowing that I was interested in microbial communities, gave me the opportunity to participate in writing an NIH grant proposal in the area of microbiome research. Writing that proposal was an education in itself, where I caught my first glimpse of the enormous work required to fund a successful lab! The grant was funded, and from that point (January 2014) until Summer 2016, we were working hard on this project. This was the most collaborative project I participated in as a graduate student, and the project with the heaviest "wet lab" component. I want to thank all the co-authors on this paper for sticking with this project through many challenges and changes.

7.2 Synopsis

The altered Schaedler flora (ASF) is a model microbial community with both *in vivo* and *in vitro* relevance. Here we provide the first characterization of the ASF community *in vitro*, independent of a murine host. We compared the functional genetic content of the ASF to wild murine metagenomes and found that the ASF functionally represents wild microbiomes better than random consortia of similar taxonomic composition. We developed a chemically-defined medium that supported growth of seven of the eight ASF members. To elucidate the metabolic capabilities of these ASF species—including potential for interactions such as cross feeding—we performed a spent media screen and analyzed the results through dynamic growth measurements and non-targeted metabolic profiling. We found that cross-feeding is relatively rare (32 of 3,570 possible cases), but is enriched between *Clostridium* ASF356 and *Parabacteroides* ASF519. We identified many cases of emergent metabolism (856 of 3,570 possible cases). These data will inform efforts to understand ASF dynamics and spatial distribution *in vivo*, to design pre- and probiotics that modulate relative abundances of ASF members, and will be essential for validating computational models of ASF metabolism. Well-characterized, experimentally tractable microbial communities enable research that can translate into more effective microbiome-targeted therapies to improve human health.

7.3 Introduction

The microbiome is enormously complex and its composition varies not only between individuals, but within the same individual spatially and temporally [1, 2]. Most of the microorganisms that comprise the microbiome variously interact through forms of competition and cooperation that are largely uncategorized [3]. Despite the daunting complexity of this system, a great deal of research effort is expended with the goal of identifying governing principles that will allow prevention and treatment of a range of human conditions connecting the immune system [4], diet and metabolism [5], emotional health [6], and other relevant systems [7] to the microbiome. If sufficiently understood, there is enormous therapeutic potential in microbiome modulation.

It is well-established that the composition of microbial communities is linked to host health, but many studies linking the microbiome to health-related outcomes provide descriptive or correlative results rather than establish causation [3, 8]. Additionally, despite growing databases of reference genomes, many species detected in these studies are new, if they are detected at all [9]. Germ-free animals colonized specifically with known microorganisms—gnotobiotic animals—enable experiments which can establish causation [10]. Such experiments cannot easily be performed in humans, making gnotobiotic animals crucial to studying microbiome structure and function.

Germ-free and gnotobiotic mice often do not develop

normal immune systems or gastrointestinal function [11]. This problem was addressed in part by work in which a cocktail of eight microbial species known as the Altered Schaedler Flora (ASF) was identified [12, 13]. Germ-free mice colonized exclusively with the ASF develop relatively normal immune systems and gastrointestinal function [14, 15]. ASF-colonized mice are commercially available and widely used [16–23]. The ASF serves as an experimentally tractable surrogate for wild-type microbiomes.

A limiting factor in ASF-based research to-date is the paucity of knowledge about the eight species contained in the ASF. Little is known about the genetics, metabolism, or *in vitro* characteristics of these eight species, because their primary value historically has been to standardize mice [13]. Draft genome sequences for all eight ASF member species were published in 2014 [24]. Future efforts to understand the mechanistic underpinnings of ASF-host interactions, or ASF dynamics within the host, will depend on a much deeper knowledge of the physiology and metabolism of each ASF member individually, and the interactions among them. To facilitate this goal, we exhaustively compared the functional gene content of all ASF species among each other, to wild-type murine metagenomes, and to random consortia of similar taxonomic composition. We developed a chemically-defined medium and performed the first *in vitro* analysis of the growth and metabolism of ASF member species. Finally, we experimentally determined the effects on growth and metabolism of spent media interactions between members of the ASF. The results of this study will serve as a resource for future ASF-based research, and provide a strong foundation for future computational modeling efforts. By better understanding the ASF—including interactions between its members—it will be possible to glean more from ASF-based mouse experiments, thus increasing the value of ASF-colonized mice as a model system for microbiome-host interactions.

7.4 Materials and Methods

Strain Information

All strains are identified by the associated ASF number. We performed experiments with ASF356 (*Clostridium* sp.), ASF360 (*Lactobacillus intestinalis*), ASF361 (*Lactobacillus murinus*), ASF457 (*Mucispirillum schaedleri*), ASF492 (*Eubacterium plexicaudatum*), ASF500 (*Pseudoflavonifractor* sp.), ASF502 (*Clostridium* sp.), and ASF519 (*Parabacteroides goldsteinii*) [15]. All strains are grown in an anaerobic chamber (Shel Lab BactronEZ, Cornelius, OR, USA) with mixed anaerobic gas (5% Carbon Dioxide, 5% Hydrogen, 90% Nitrogen) at 37 °C. Anaerobic conditions were confirmed periodically using an anaerobic indicator (Oxoid, Basingstoke, UK). All strains were propagated on supplemented Brain-Heart Infusion (BHI) agar.

Media Preparation

Supplemented BHI medium: Brain-Heart Infusion base (BD, Franklin Lakes, NJ, USA) was supplemented with yeast extract, hemin (0.005 g/l), L-cysteine (0.25 g/l), vitamin K1 (9.84 mg/l) and 5% each of newborn calf serum, horse serum, and sheep serum.

Supplemented LB medium: LB base in powder form (Sigma, St. Louis, MO, USA) was combined with L-cysteine (Sigma), added KH_2PO_4 (6 g/l), $(NH_4)_2SO_4$ (6 g/l), NaCl (12 g/l), $MgSO_4{\cdot}7H_2O$ (2.5 g/l), $CaCl_2{\cdot}2H_2O$ (1.6 g/l), L-cysteine (0.25 g/l) (see detailed formulation in the Supplemental Materials) and deionized water, which was autoclaved at 121 °C for 20 min. After cooling, vitamin K1 (9.84 mg/l) and filter sterilized (0.22 μm pore size) solutions of hemin (0.005 g/l), lactose (0.05 g/l), and Tween-20 (0.01 g/l) were added.

All media was equilibrated overnight in the anaerobic chamber before inoculation with ASF members.

Genomic Analysis and Comparison with Wild Murine Microbiota

Shotgun sequencing metagenomic data from the feces of 15 wild mice from a previous study [25] were used as a reference data set (Shannon diversity of 163 $\pm$ 72) for comparative analysis to the ASF. We downloaded protein sequences for all 15 samples, which were then annotated with HMMER Version 3.1b2 [26], using bactNOG (144,498 protein sequences) from eggNOG version 4.1 [27] as the profile hidden Markov models. For each gene call, a non-supervised orthologous group (NOG) was assigned using the database target with the lowest e-value below 10^{-10}. Overall, 22.3% of metagenomic open reading frames were assigned a NOG annotation. Protein sequences for each ASF species were downloaded from GenBank and annotated using the same procedure.

To compare metagenome coverage by the ASF to coverage by random communities, species were drawn from among the 989 Firmicutes and 176 Bacteroidetes in bactNOG in a 6:2 ratio, respectively, to represent the most abundant phyla in the mouse gastrointestinal tract [28]. The complete list of Firmicutes and Bacteroidetes in bactNOG is available in Supplemental Data 1. Random communities of size 8, 16 and 32 were compared to the ASF for percent coverage of NOGs annotated in any metagenomic sample. This coverage was further sorted by sample frequency, where each NOG can occur in up to

15 metagenomic samples. NOGs containing functional annotations in more than one category were discarded during all portions of analysis (representing <1% of total annotations in any sample).

Preparation of Spent Media

Spent media from each ASF member was prepared by growing each species in supplemented LB for 70 h. The resulting culture was centrifuged at 3,500 rpm for 10 minutes and the supernatant was filter sterilized (PVDF membranes with 0.22 μm pore size). Aliquots of spent medium were stored at -80 °C. Individual aliquots were thawed and equilibrated in the anaerobic chamber overnight before inoculation.

Growth Measurements

Growth curves were obtained for ASF members in the anaerobic chamber using four miniaturized plate readers measuring optical density at 870 nm [29]. Overnight liquid cultures of 10 ml were prepared for each ASF member: The entire volume of the overnight cultures were centrifuged at 8 000 rpm for 2 minutes and the resulting pellets were resuspended in fresh liquid medium to produce a dense suspension of 0.75 ml. The optical density of the suspension was obtained on a Tecan (Männedorf, Switzerland) plate reader at 600 nm. Liquid cultures were prepared in six-well plates with 6 ml per well, and inoculated (from the dense suspension) to a starting OD600 of 0.001. Each experimental condition was replicated four times. Each plate was covered with a Breath-Easy membrane (Sigma). The OD870 was tracked for 70 h. At the final time point, the OD600 was measured for each well on the Tecan. The growth curves obtained at OD870 were normalized to the initial and final OD600 measurements (see Supplemental Materials and Methods). For each well of the 6-well plate, growth curves from four independent LED pairs were averaged to produce a single growth curve per well. To determine the area under a growth curve (AUC), we applied trapezoidal numerical integration. The R (The R Foundation, Vienna, Austria) code for growth curve analysis is available in an open online repository (see Code and Data Availability).

Determining Substrate Utilization and Byproduct Consumption with NMR Spectroscopy

Media (fresh or spent) samples of 2 ml were filter sterilized (0.22 μm pore size) and frozen at -80 °C. Standard one-dimensional (1D) 1H-NMR spectra with water pre-saturation were acquired at 300 K using a 600 MHz Avance III spectrometer (Bruker, Rheinstetten, Germany). Spectra were imported into Matlab R2014a (The Mathworks, Inc., Natick, MA, USA). Biologically irrelevant regions of the spectra were removed (TSP resonance at δ1H 0 and residual water peak δ1H 4.5-5.2) before peak alignment by recursive segment-wise peak alignment [30]. The loadings of pairwise principal component analysis models, comparing blank media with the spent media of each bacteria species, were used to identify metabolites generated or consumed in each experiment. The relevant regions of the spectra were integrated to calculate relative spectral intensities for each metabolite. Relative intensities in spent and double spent media were converted to z-scores with respect to metabolite abundances in fresh media. We defined significant abundance changes as those of magnitude greater than ± 2 standard deviations from zero (zero being the metabolite abundance in fresh media). The peak integral data and associated R code for analysis and visualization are available in an open online repository. Instances of emergent metabolism were classified by comparing metabolite presence/absence calls between single and double spent media conditions. We describe our method in the Supplemental Materials and Methods. The custom R script used to classify cases of emergent metabolism is also available in the online repository (see Code and Data Availability).

Genetic and Metabolic Similarity Analysis

We used the Jaccard distance (1 Jaccard similarity coefficient) to quantify the distance between NOG annotation sets for all pairs of ASF members. We converted the metabolomics profiles (all 85 metabolites) for each species to lists of metabolites which were consumed (z-score <-2) or produced (z-score >2) and calculated the Jaccard distance between all pairs of spent media profiles. The Python script used for this analysis is available in an online repository (see Code and Data Availability).

Code and Data Availability

Detailed methods for scanning electron microscopy, colony imaging, media preparation, and NMR metabolic profiling can be found in the Supplemental Materials. Our data and analysis scripts are available at the following repository: mbi2gs.github.io/asf_characterization/ Some large analysis output files and annotation files for metagenomic data are excluded due to file hosting size limitations, but are available upon request from the authors or can be generated using the indicated raw data, HMMer, associated eggNOG files, and scripts in the repository.

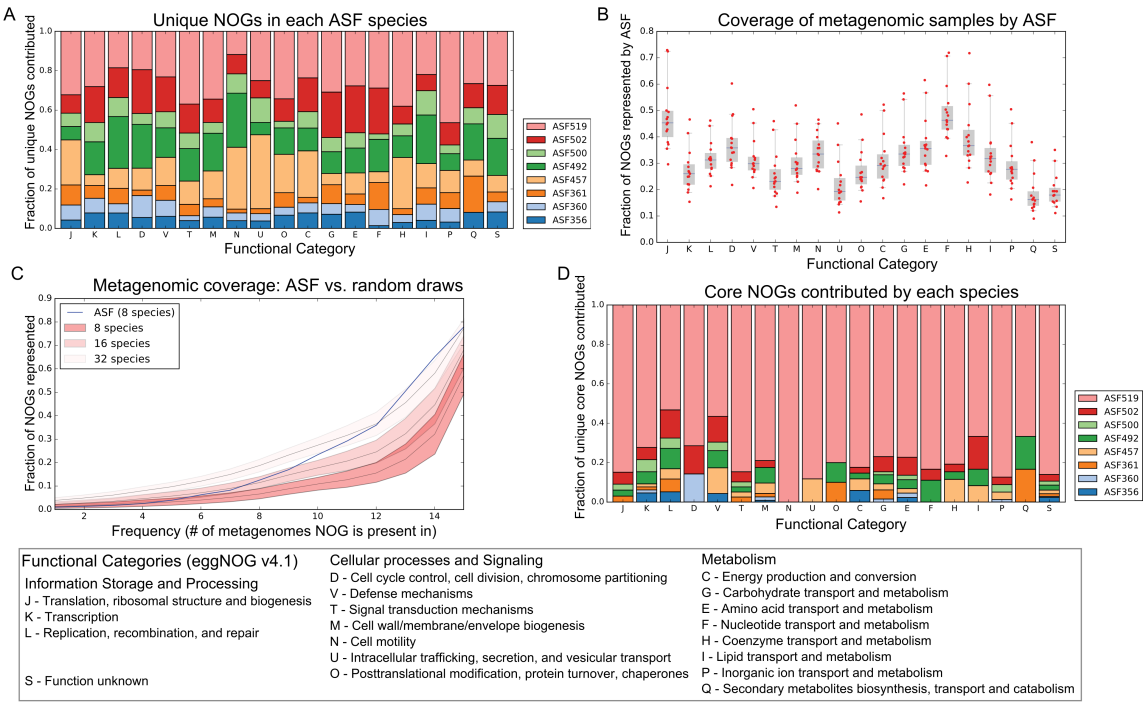


Figure 7.1: **Comparative analysis of the ASF and wild microbiomes**. A) The unique contribution of each ASF species to the ASF metagenome is relatively evenly distributed, with the unique contribution of each species being roughly proportional to genome size. Unique NOGs are those present in only 1 ASF species. B) Coverage of the 15 wild mouse fecal metagenomes by the ASF divided by NOG functional category. Coverage indicates how representative the ASF metagenome is of wild mouse metagenomes. Coverage of individual metagenomic samples is represented by red circles, median coverage is shown as a blue line within boxes, boxes extend to mean $\pm$ 1 standard deviation, and whiskers extend to 5th and 95th percentiles. Across all categories, the ASF overlaps with $\sim$35% metagenomic NOGs. C) Coverage of metagenomic NOGs by the ASF and random microbial consortia. Random consortia mimic the phylum-level distribution of the most abundant species in the mouse gastrointestinal tract. The x-axis indicates the number of metagenomes in which the NOGs are present. Coverage of metagenomic NOGs by random consortia of 8, 16, and 32 species (dark to light shading, respectively) are indicated as median lines surrounded by 5th/95th percentile distributions. The ASF covers core metagenomic NOGs (core NOGs occur in all 15 samples) better than any combination of 8 or 16 species and better than the median of 32 species. D) Unique contribution of each ASF species to core metagenomic NOGs. *Parabacteroides* ASF519 contributes the majority of core NOGs in every category.

7.5 Results

Development of a Defined Medium

We developed a growth medium with defined chemical composition that supports the anaerobic growth of all ASF members (excluding *Mucispirillum* ASF457). This novel, defined medium is based on standard LB medium, supplemented with minerals, salts, and components commonly added to support growth of anaerobes (see detailed formulation in the Supplemental Materials). LB is not generally considered a "chemically defined" medium because of complex ingredients such as yeast extract. However, previous research has identified the components of LB to a degree suitable for computational metabolic models and metabolomics purposes [31, 32]. We confirmed the presence of the majority of expected metabolites using NMR spectroscopy (Supplemental Table 1).

Morphology and Appearance

We describe the cellular and colony morphologies of all eight ASF members in the supplemental materials (Supplemental Figures 1 and 2).

ASF Compared to "Wild-Type" Murine Microbiome

We annotated the ASF genomes using the eggNOG database and identified unique genetic content within each species (Figure 7.1A). Non-supervised Orthologous Groups (NOGs) are clusters of highly similar proteins, where proteins within each cluster generally share the same function. We found that all eight species possess

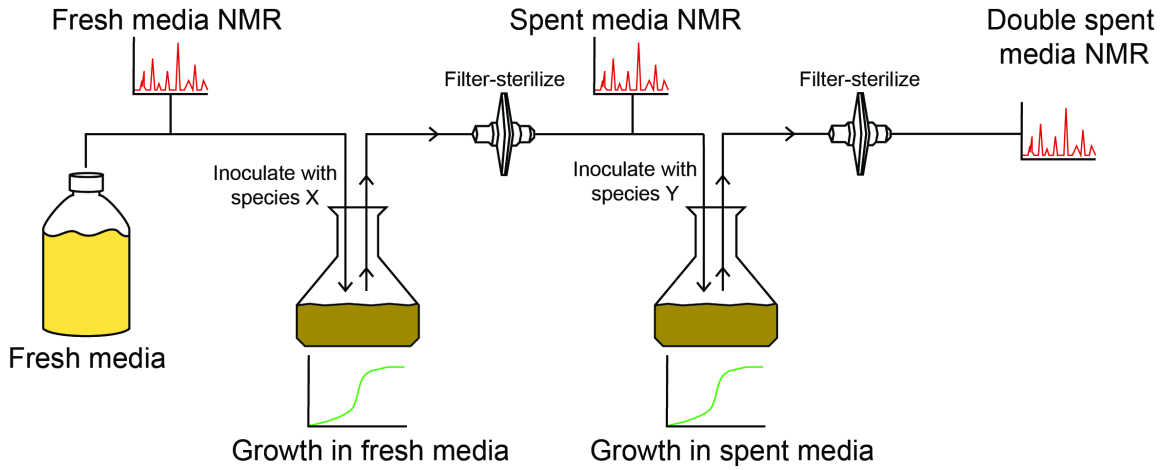


Figure 7.2: **Spent media experimental setup**. Each ASF member species was grown independently in fresh growth medium. Growth was monitored for 70 h by optical density, which allowed for comparison of growth rates between species. The supernatant from these first cultures was filter sterilized to produce "spent media", which was subsequently profiled by NMR spectroscopy and compared to the fresh growth medium. This initial metabolomics analysis identified the metabolites utilized and byproducts produced by each species. For the second round, each ASF species was inoculated into spent medium from the other species. Growth was monitored and compared to growth in fresh medium. The supernatant from this second round ("double spent media") was filter sterilized and compared to the spent medium from which it originated to identify further metabolites that were used or consumed.

unique NOGs in proportion to genome size [24]. We next compared the composite metagenome of the ASF to the NOGs found in 15 wild murine metagenomes (Figure 7.1B). We found that the composite ASF metagenome overlaps with the murine microbial metagenome by ∼35% in each functional category. Given that the ASF was developed specifically as a surrogate murine microbiome, we hypothesized that the ASF would share key functions with wild type microbiomes; functions which would be less common in random microbial consortia. We compared the composite ASF metagenome to 10 000 random microbial consortia with similar taxonomic composition (Figure 7.1C). We sorted NOGs by sample frequency (i.e. presence in 1–15 metagenomic samples; frequency distribution shown in Supplemental Figure 3) and determined a core group of NOGs that occurred in all 15 wild murine metagenomic samples (3,611 NOGs out of 135,013 unique NOGs observed). Surprisingly, the ASF shares more gene content with the wild metagenomes than any random 8-species consortia. Larger random consortia approach (16 species) and exceed (32 species) the ASF coverage of the wild metagenomes. However, the ASF maintains better coverage of core NOGs (those that occur in all 15 wild microbiomes) than any 16-species consortia and the median of 32-species consortia. Additionally, we found that replacing a Bacteroidetes in the random consortia with *Mucispirillum* ASF457 (a member of the phylum Deferribacteres) decreased the coverage of core NOGs (Supplemental Figure 4), demonstrating that the Deferribacteres phylum does not explain the superior coverage by the ASF. The ASF contains 2,820 of the 3,611 (78.09%) core NOGs. Of these 2,820 core NOGs, 1 283 (45.50%) are unique to a single ASF species. Of these unique NOGs 1,036 (80.75%) are contributed by *Parabacteroides* ASF519, representing the majority of unique core NOGs in every functional category (Figure 7.1D). These findings suggest that *Parabacteroides* ASF519 is primarily responsible for the ASFs high coverage of the core wild murine metagenome.

Individual Growth Characteristics

Each ASF member grew (excluding *Mucispirillum* ASF457) in fresh supplemented LB medium (Supplemental Figure 5). *Lactobacillus* ASF361, *Parabacteroides* ASF519 and *Clostridium* ASF356 grew most rapidly, while *Pseudoflavonifractor* ASF500, *Eubacterium* ASF492 and *Clostridium* ASF502 grew most slowly. Taxonomic relatedness did not necessarily predict growth rates well, given that *Clostridium* ASF356 (a fast grower) and *Clostridium* ASF502 (a slow grower) are both members of the genus *Clostridia*. Similarly, the two *Lactobacilli*, *Lactobacillus* ASF360 and *Lactobacillus* ASF361, vary drastically in growth rate—*Lactobacillus* ASF361 grew more quickly and to a higher density in liquid and on solid media.

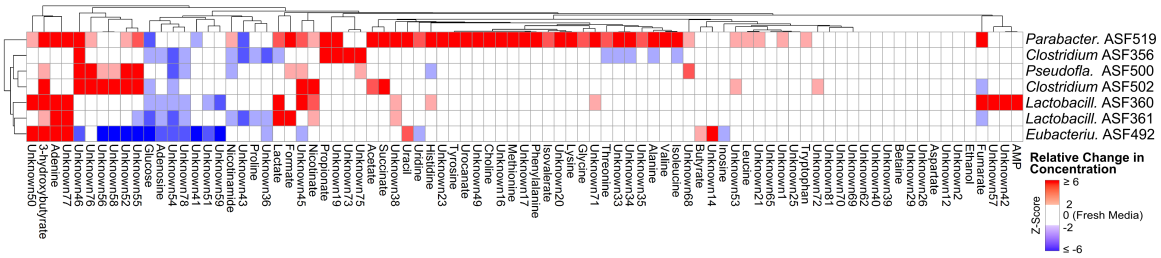


Figure 7.3: **Relative changes for 85 NMR peaks in single spent media samples**.NMR peak integrals are proportional to metabolite concentrations. Relative changes in peak integrals are displayed as z-scores relative to fresh media, with zero (white) indicating that the metabolite concentration is the same as in fresh media, while values greater than 2 standard deviations above (red) or below (blue) fresh media indicates higher or lower concentrations than fresh, respectively. Z-scores $\leq$ -6 or $\geq$ 6 are displayed as -6 or 6, respectively. Rows correspond to individual ASF members. For example, the first row indicates the metabolite z-scores relative to fresh media after the growth of *Parabacteroides* ASF519.

Interactions Characterized Using Spent Media

We characterized directional, species-species interactions by screening all pairs of ASF members through a series of spent media experiments (Figure 7.2). In brief, spent medium was prepared for each species by growing it in fresh liquid media for 70 h. We use the notation "spentXXX" to indicate the supernatant resulting from growth of ASFXXX (e.g. "spent356" to indicate the spent media resulting from growth of *Clostridium* ASF356). Having reached stationary phase, the supernatant from the culture was filter sterilized. This resulting spent medium was used to culture each ASF member in turn. The loss of some substrates and addition of new byproducts from the first species influenced the growth of subsequent species. The supernatant resulting from the growth of a second species in the spent media from a previous species is referred to as "double spent media".

Growth Inhibition

No species was able to grow in its own spent medium, which is consistent with the expectation that a species has exhausted a media environment once it has entered stationary phase (Supplemental Figure 5). The majority of interactions resulted in decreased growth or completely stifled growth in the second species. *Parabacteroides* ASF519 was able to grow in the spent media from most other member species with the exception of spent medium from *Lactobacillus* ASF361. *Lactobacillus* ASF360 and *Pseudoflavonifractor* ASF500 did not grow in spent media from other species. Marginal growth was observed with *Lactobacillus* ASF361 grown on spent492 (i.e. spent media produced by *Eubacterium* ASF492). *Lactobacillus* ASF361 prevented growth of all other members, while spent media from *Clostridium* ASF356 and *Parabacteroides* ASF519 prevented growth of all species with the exception of each other.

Metabolic Profiling of Spent Media

NMR spectra were obtained for all fresh, spent, and double spent media conditions (Supplemental Figure 6 and Supplemental Metabolomics Plots). Across all samples, 85 NMR peaks exhibited significant variation (Supplemental Figure 6). We were able to confidently map 36 peaks to known metabolites in our library of reference spectra.

Significant metabolic differences were observed among the ASF members growing in fresh media (Figure 7.3). *Parabacteroides* ASF519 consumed the fewest metabolites (only glucose, Unknown41 and Unknown43), while it produced many other metabolites including amino acids (alanine, glycine, histidine, isoleucine, leucine, lysine, methionine, phenylalanine, threonine, tryptophan, tyrosine, and valine). *Clostridium* ASF356 uniquely consumed isoleucine, valine, alanine, threonine and lactate and some unidentified metabolites (Unknowns 33, 34, and 35). *Eubacterium* ASF492 was the only species to consume uridine and several unidentified metabolites (Unknowns 45, 46, 56, and 58), while *Pseudoflavonifractor* ASF500 was the only consumer of histidine. *Eubacterium* ASF492 was the only species to produce butyrate in fresh media. Glucose was clearly consumed by all species except *Pseudoflavonifractor* ASF500. Nicotinamide, adenosine and two unidentified metabolites (Unknowns 54, 78) were also consumed by most species. While all spent media was acidic, *Lactobacillus* ASF361 produced the most acidic spent media (Supplemental Table 2). The reproducibility of these NMR-based observations was confirmed by comparison with an independent set of biological replicates and subsequent NMR metabolomic profiling (Supplemental Metabolomics Plots).

We interpreted the double spent samples by comparing them to the spent media from which they were derived (Figures 7.4A and 7.4B). If a species was able to

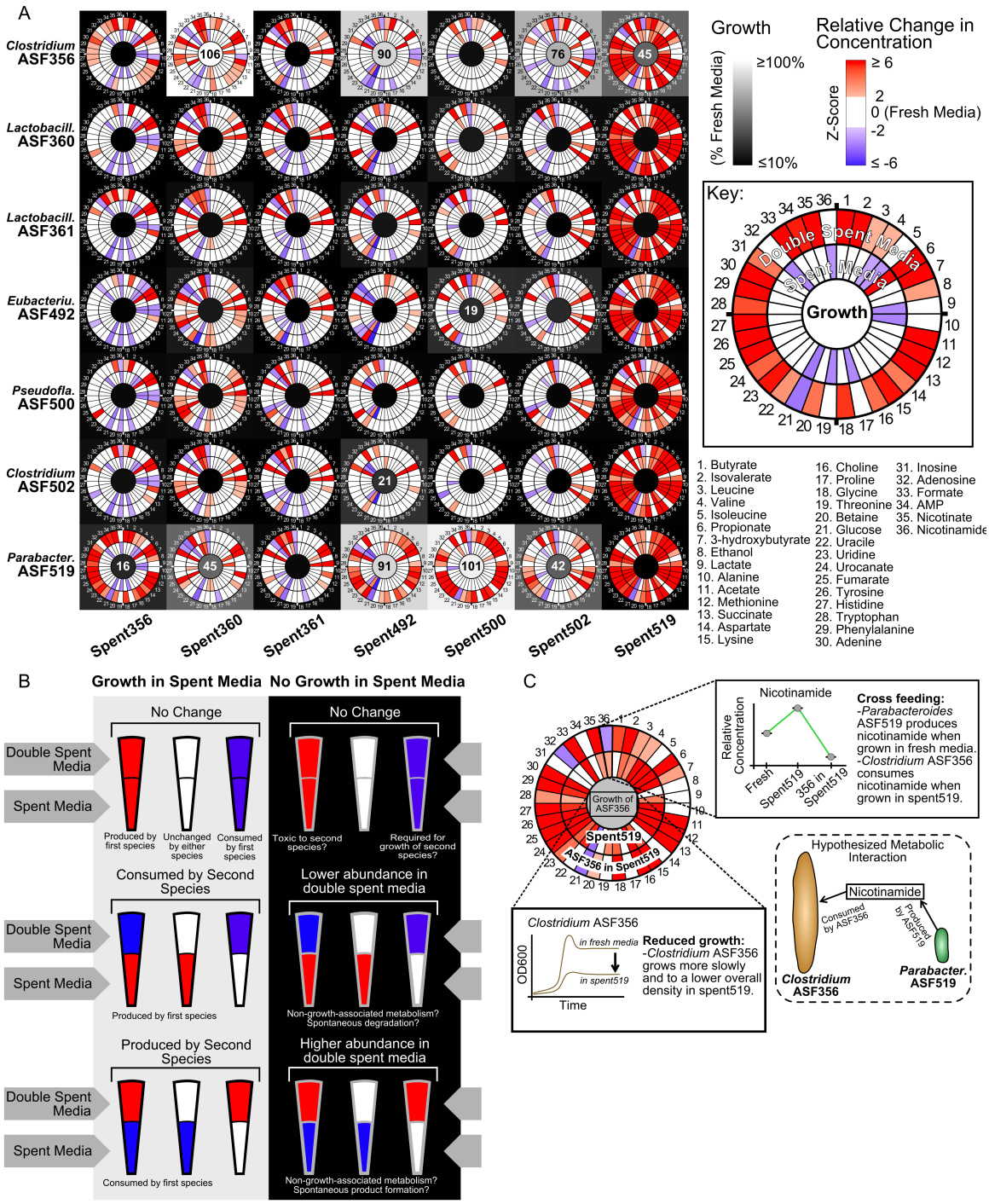


Figure 7.4: **Metabolomics analysis of all media conditions**. A) The black and white heat map underlying the figure indicates the growth achieved under spent media conditions compared to fresh media conditions. Growth is quantified using the area under the curve (AUC), indicated as a percentage of the AUC when a species is grown in fresh media. White indicates growth equal in rate and density to fresh media conditions (e.g. *Parabacteroides* ASF519 grown in spent500), while black indicates complete inhibition of growth (e.g. *Parabacteroides* ASF519 grown in spent361). For species that achieved an AUC of at least 10%, we annotate the AUC in the center of the appropriate tile. Circular heat maps within each cell display the metabolomics profiles for the spent media in that column (inner ring) and the "double spent" media (the result of growing the species from that row in the spent media from that column; outer ring). Metabolite concentrations are quantified as z-scores relative to fresh media, and are displayed as circular heat maps. Zero (white) indicating that the metabolite concentration is the same as in fresh media, while values greater than 2 standard deviations above (red) or below (blue) fresh media indicates higher or lower concentrations than fresh, respectively. B) Qualitatively, there are 18 possible scenarios when comparing double spent media to the spent media from which it was derived. In general, a metabolite can increase, decrease, or remain the same, and the interpretation of that behavior is related to whether there was observed growth in that condition. For example, a metabolite that is depleted by the first species and remains so (no change) under a no growth condition may indicate a metabolite which was required for growth of the second species. Alternatively, if a metabolite is produced by the first species, consumed by the second and growth is observed, this constitutes evidence for cross feeding. C) An example: Nicotinamide is elevated in spent519 (inner ring) and depleted when *Clostridium* ASF356 is grown in spent519 (outer ring). *Clostridium* ASF356 does grow (AUC=45%), so we hypothesize that in a co-culture, *Clostridium* ASF356 would benefit from *Parabacteroides* ASF519 producing nicotinamide.

grow in a spent media, the associated changes in metabolite abundances can be attributed to metabolic activity of that species. Of interest are those metabolites which may contribute to cross-feeding or competition in a co-culture setting. If growth is inhibited in a spent media, the metabolite profiles of the spent media can indicate compounds required for growth, or alternatively, toxic compounds.

Of the 3 570 metabolite comparisons between spent and double spent media conditions, 2 695 (75%) were unchanged between the spent and double spent conditions (Table 7.1). Of these, 2 081 (77%) unchanged metabolites were associated with conditions where the second species did not grow. Of the 2 550 comparisons where the second species did not grow, 469 (18%) changed. If a metabolite changed between spent and double spent conditions, usually it increased (717 instances of 875 changed, or 82%).

Cases of potential cross feeding were rare, where the second species grew and simultaneously consumed a metabolite produced by the first species (32 instances of 875 changed, or 4%). *Clostridium* ASF356 was able to grow in spent519 (AUC=45% of fresh media growth; Figure 7.4C) and was the condition most enriched for cases of potential cross-feeding (10 of 85 possible metabolites, or 12%). As an example of potential cross-feeding, *Parabacteroides* ASF519 produced nicotinamide when grown in fresh media, and *Clostridium* ASF356 appears to have consumed the nicotinamide in the spent519 media (Figure 7.4A and C). Given these data, we hypothesize that cross-feeding may occur in a co-culture setting such that *Clostridium* ASF356 would consume nicotinamide produced by *Parabacteroides* ASF519. Between spent519 and *Clostridium* ASF356 grown in spent519, similar cross-feeding-like profiles are observed for alanine, isoleucine, lactate, threonine, uracil and several unidentified metabolites (Unknowns 35, 52, 55, and 76).

There were 856 instances of emergent metabolism (of 3,570 possible), 168 of which occurred in only one condition. For instance, we observed cases of emergent biosynthesis, such that a species produced a given metabolite only when grown in the spent media of another species. *Parabacteroides* ASF519 produced butyrate, betaine, and several unidentified metabolites (Unknowns 2, 12, 14, 36, 39, and 40), but only when grown on spent356 or spent500 (Supplemental Figure 6). There were several cases where the emergent phenotype occurred in a single condition. For example, *Clostridium* ASF356 only produced aspartate, lysine, methionine, phenylalanine, succinate, tyrosine and several unidentified metabolites (Unknowns 23, 25, 26, and 65) when grown in spent360. Alternatively, while *Parabacteroides*

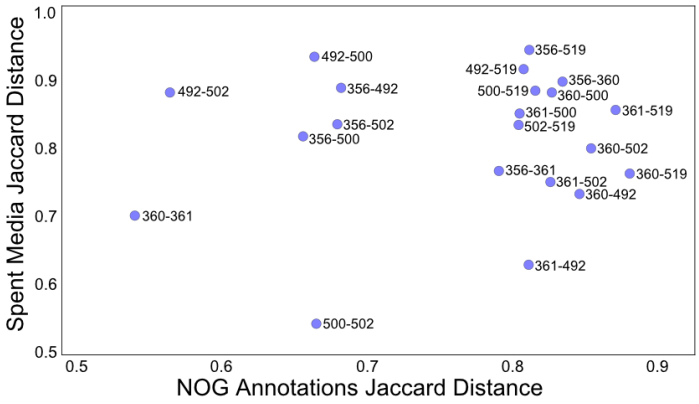


Figure 7.5: **Genetic distance associated with greater variance in metabolic distance**.We quantified the genetic distance between all species pairs using the Jaccard distance between the NOG annotation sets. We similarly quantified the distance between the spent media metabolomics profiles for all pairs of ASF members. Genetic similarity is not strongly correlated with metabolic state under these conditions. Points are labeled with the ASF identifiers for the species pair.

ASF519 produced 3-hydroxybutyrate in every other media condition, when grown in spent502 it switched to consuming 3-hydroxybutyrate. The complete list of emergent, metabolic observations is available in Supplemental Data 2.

Genetic Distance Associated with Variance in Metabolic Distance

We quantified the genetic distance between all pairs of ASF members using the Jaccard distance between the NOG annotations for each pair. Using the same metric, we quantified the distance between spent media metabolomics profiles for all pairs of ASF members. We found that genetic similarity is not strongly correlated with metabolic state under these conditions (Figure 5). Indeed, some pairs of closely related species (e.g. *Lactobacillus* ASF360 and *Lactobacillus* ASF361) were more different in terms of spent media profiles than some more distant pairs (e.g. *Pseudoflavonifractor* ASF500 and *Clostridium* ASF502). Furthermore, we performed a correlation analysis and identified 11,079 NOG-metabolite pairs which were statistically significant (Spearmans correlation and Bonferonni multiple testing correction with n=160 746; Supplemental Figure 7). After excluding unique NOGs and metabolites which were consumed or produced by a single species, 458 correlations were significant.

	Growth			Subtotal	No Growth			Subtotal	Totals
No change	High	Medium	Low		High	Medium	Low		
	146	416	52	614	460	1,426	195	2,081	**2,695**
Lower in Double Spent	High to Low	High to Med.	Med. to Low		High to Low	High to Med.	Med. to Low		
	2	30	41	73	3	43	39	85	**158**
Higher in Double Spent	Low to High	Low to Med.	Med. to High		Low to High	Low to Med.	Med. to High		
	9	51	273	333	10	43	39	384	**717**

Table 7.1: **Classification of metabolite profiles between spent media and double spent media (known metabolites only)**. Categories correspond to those presented in Figure 7.4B. Metabolite relative abundance indications: "High" indicates a relative abundance 2 standard deviations above that in fresh media; "Med." indicates an abundance within ±2 standard deviations of that in fresh media; "Low" indicates a relative abundance 2 standard deviations below that in fresh media. For example, where a second species grew in the spent media of a first species ("Growth" column), there were 9 cases where a metabolite which was "low" in the spent media increased to "high" in the double spent media ("Low to High").

7.6 Discussion

We present a novel approach to characterizing microbial communities *in vitro*, and the results of applying this approach to gain insights into a model microbial community known as the ASF. Through a bioinformatics analysis, we found that the ASF is far more representative of wild microbiome functionality than random consortia of similar or larger size. Through a spent media screen, we found that cross-feeding interactions are relatively rare, while non-growth-associated and emergent metabolism are relatively common. These, together with the rest of our findings, form the beginnings of a rich knowledge base which increases the utility of the ASF as a model gut community.

A primary outcome of this work is standardization of ASF resources. Firstly, our morphological descriptions accompanied by images, all gathered under the same conditions, provide a reference for future researchers. Comparing morphology to references such as these will improve reproducibility and support the discovery of new phenotypes. Secondly, we developed a chemically-defined LB-based medium which simplifies metabolic profiling, and ongoing efforts to build genome-scale metabolic network reconstructions for the ASF members. One drawback of the new LB-based medium is that it does not support the growth of *Mucispirillum* ASF457 (*M. schaedleri*), which is difficult to grow effectively, even in complex media [16]. We attempted, unsuccessfully, to identify media components which would allow *Mucispirillum* ASF457 to grow, including the addition of porcine mucin, given the fact that *Mucispirillum* ASF457 colonizes the mucous layer in the murine colon [33]. We also attempted to grow *Mucispirillum* ASF457 in the spent media of other ASF members. This is an area for future research (Supplemental Data 3 includes a list of gene annotations missing from *Mucispirillum* ASF457 that are present in all other ASF members). Despite this shortcoming, the new media successfully enabled metabolic profiling of the remaining seven ASF members and their interactions.

The presented genomic analysis of the ASF vastly expands our knowledge of how the ASF relates—on a functional level—to more complex microbiomes. Indeed, the ASF is far simpler than a wild-type microbiome in terms of both species and genetic composition. We found that *Parabacteroides* ASF519 is a major contributor to the unique qualities of the ASF, with impressive coverage of genes and metabolic activities that may be vital in the wild mouse microbiome. Additional studies are needed, which reach beyond coverage of functional orthologs, to better understand both the essential and redundant roles played by each ASF species.

Traditional co-culture experiments have several drawbacks which make it challenging to determine the mechanism underlying interactions between two species. These include difficulties determining which species utilized or produced a given metabolite [34], and measuring growth of individual species which requires tools with lower temporal resolution and much higher costs than optical density [35–38]. Computationally inferring interactions from metagenomic data generally cannot resolve interaction directionality, but rather is limited to identifying correlations between species abundances [39]. Spent media experiments resolve some issues confronted in co-culture experiments and computational inference. By separating interactions into two steps (Figure 7.2), it is simple to infer interaction directionality, and straightforward to generate hypotheses about underlying mechanisms [40, 41]. Growth measurements can be gathered at high resolution with metrics such as optical density because only a single species is growing. It should be noted, that spent media experiments do not allow for cell-to-cell contact or dynamic signaling between species,

which may otherwise be relevant in co-culture or *in vivo* [42]. Moreover, the nature of interactions can be context-specific, such that interactions identified in this spent media screen are not definitive for all conditions [43]. Finally, while we utilized 100% spent media in this study, we expect that a wider range of growth phenotypes will be observable by using spent media dilutions or by creating "partially spent media" such that there are sufficient nutrients to support growth, but molecules from the first species would still be able to influence the second species. This is a promising direction for future research.

Our interpretation of these spent media experiments relies first on comparing the growth dynamics of a given species in both fresh media and spent media (Supplemental Figure 5). Subsequent metabolic profiling (Figure 7.4) identified specific compounds hypothesized to play a role in causing the observed interaction dynamics. In a similar spent media experiment between environmental isolates of leaf degrading bacteria, it was found that natural isolates engage in less cross-feeding than isolates evolved together for several generations in vitro [41]. While the ASF has evolved for many generations as a single community within mice, the defined media environment used in our *in vitro* experiments is very different from the murine gut environment, and so it is not surprising that we observed few instances of cross-feeding. An interesting future direction would be to evolve the ASF *in vitro* in the defined medium, after which we would predict that more cross-feeding would be observable. Our observation that emergent metabolic phenotypes are common between ASF members (at least one emergent phenotype identified between all pairwise species interactions) agrees with recent computational and *in vitro* work demonstrating that the vast majority of microbial pairs and media conditions exhibit emergent biosynthetic behaviors [44, 45]. Also of interest, we observed 469 instances in which metabolite relative abundances changed despite an absence of growth (e.g. *Lactobacillus* ASF361 grown in spent500). One explanation is that in these cases, inoculated cells are metabolically active without active cell division [46]. It is notable that most species grew slower (or not at all) and to a lower overall density in spent media (Figure7.4 and Supplemental Figure 8). This is likely due to the rich media conditions, in that it is unlikely that another species could produce a mixture of compounds more growth-promoting than already found in the fresh media. We would expect to see more growth-enhancing interactions under minimal media conditions.

We found that the similarity of genetic annotations between any pair of ASF members was not strongly correlated with the metabolic phenotypes of those same pairs (Figure 7.5). This result is not surprising given that the relationship between the phylogenetic distance and the metabolic capabilities of two species can be modeled by an exponential , the relationship is neither linear nor strong [47]. Furthermore, we identified several cases where the NOG presence/absence distribution was significantly correlated with the consumption or production of specific metabolites (Supplemental Figure 7). However, because there are so few species, the vast majority of these significant correlations are between NOGs found in a single species and the metabolites which were uniquely consumed or produced by that species. These results reinforce the need for more sophisticated approaches to linking genotype to phenotype, for example, using comparative metabolic network modeling [48].

Of the seven ASF members that grew in the supplemented LB medium, *Clostridium* ASF356 and *Parabacteroides* ASF519 grew most rapidly and to the highest overall density (Supplemental Figure 5). Furthermore, both species grow in spent media from each other (Figure 7.4). A summary of the data for *Clostridium* ASF356 and *Parabacteroides* ASF519 highlights competition for glucose and many opportunities for cross-feeding (Figure 7.6). *Clostridium* ASF356 consumed a far more diverse assortment of metabolites, while *Parabacteroides* ASF519 consumed very few. *Parabacteroides* ASF519 has a more diverse metabolic output than *Clostridium* ASF356. Considering ASF spatial distribution in vivo, *Clostridium* ASF356 is more abundant in the cecum and *Parabacteroides* ASF519 is the most abundant ASF member in the colon [16]. *Parabacteroides* ASF519 appears to be a scavenger, growing robustly in the distal colon where the ability to produce essential biomass components from few inputs is an advantage.

The observation that *Parabacteroides* ASF519 required few substrates is partially explained by its large genome size (6.87 Mb), the largest of the ASF. We would expect large genome size to correlate with greater biosynthetic capacity, knowing that smaller genomes correlate with auxotrophy [49]. The metabolic characteristics of *Parabacteroides* ASF519 are also interesting in light of our comparison of random consortia to the ASF: the core functions found in fecal metagenomes were covered best by *Parabacteroides* ASF519 (Figure 7.1D), which allowed the ASF to out-perform much larger microbial consortia (Figure 7.1C). Naturally, the large genome of *Parabacteroides* ASF519 could lead to more coverage of the core metagenome. However, the next two largest genomes from *Eubacterium* ASF492 (6.51 Mb) and *Clostridium* ASF502 (6.48 Mb) do not come close to the same coverage, nor do these species display the same prolific biosynthetic activity under these

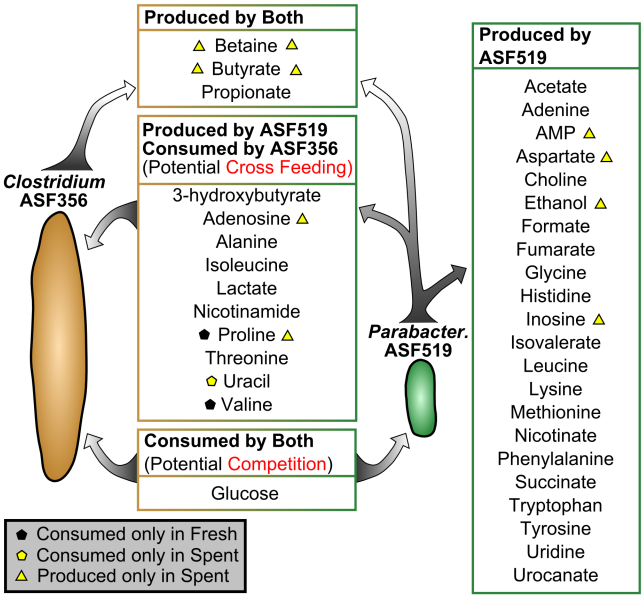


Figure 7.6: **Inferred metabolic interactions between *Clostridium* ASF356 and *Parabacteroides* ASF519.** *Clostridium* ASF356 and *Parabacteroides* ASF519 were able to grow in many more spent media conditions than other species, including spent media from each other. We combined evidence from four media conditions including spent356, spent519, *Clostridium* ASF356 grown in spent519 and *Parabacteroides* ASF519 grown in spent356 to form our hypothesis of the metabolic interactions that would occur in co-culture. Black pentagons indicate metabolites that are consumed only from fresh media, yellow pentagons indicate metabolites that are consumed only from spent media, and yellow triangles indicate metabolites that are produced only in spent media. In general, *Parabacteroides* ASF519 produces many more compounds than *Clostridium* ASF356, while *Clostridium* ASF356 consumes many more compounds than *Parabacteroides* ASF519. Both species consume—and would be expected to compete for—glucose. Both species produce propionate in abundance, while both species also produce butyrate and betaine, but only when grown in the spent media from the other. For clarity, we have excluded unidentified NMR peaks from this figure.

in vitro conditions. *Parabacteroides* ASF519 is an unexpectedly vital contributor to the ASF metagenome and metabolic activity.

Our analysis of the ASF metagenome and spent media experiments have produced a profile of ASF genetics and metabolism which will enable future research to leverage knowledge of the unique qualities of the ASF. As an example of how the ASF can be used advantageously, a recent study leveraged an understanding of ASF metabolism to engineer a microbiome which improved survival of mice with hepatic injury [20]. An example where more detailed information about ASF metabolism would have been highly relevant, a recent study colonized gnotobiotic mice with a subset of the ASF in an attempt to prevent butyrate production [50], knowing that the full ASF community does produce butyrate *in vivo* [51]. That subset correctly excluded *Eubacterium* ASF492, but included *Parabacteroides* ASF519, which we found to also produce butyrate. Future experiments excluding *Parabacteroides* ASF519 from ASF-colonized mice will enhance our understanding of its impact on mouse health and metabolism, and could shed light on the role of similar species in the gastrointestinal tract of humans. The metabolomics profiles presented in this study indicate nutritional supplements which could be used as pre-biotics in ASF-colonized mice. Indeed, greater understanding of the ASF increases its utility as a testing ground for validating strategies for the development of microbiome-targeted therapies.

The ASF is a unique microbial community with a long history of use in murine models, with untapped potential to become a highly characterized model microbiome. Our characterization of ASF morphology, functional genetic content, growth, metabolism and interactions lays a strong foundation for future research into gut ecology and efforts to engineer the gut microbiome to improve health

7.7 Acknowledgements

The authors thank Michael Wannemuehler and Gregory Phillips at Iowa State University for providing strains of all eight ASF members.

7.8 References

[1] Lozupone CA et al. “Diversity, stability and resilience of the human gut microbiota”. In: *Nature* 489.7415 (Sept. 2012), pp. 220–230. DOI: 10.1038/nature11550.

[2] Faith JJ et al. “The Long-Term Stability of the Human Gut Microbiota”. In: *Science* 341.6141 (July 2013), pp. 1237439–1237439. DOI: 10.1126/science.1237439.

[3] Ji B and Nielsen J. “From next-generation sequencing to systematic modeling of the gut microbiome”. In: *Frontiers in Genetics* 6 (June 2015). DOI: 10.3389/fgene.2015.00219.

[4] Gevers D et al. “The Treatment-Naive Microbiome in New-Onset Crohn’s Disease”. In: *Cell Host & Microbe* 15.3 (Mar. 2014), pp. 382–392. DOI: 10.1016/j.chom.2014.02.005.

[5] Wang J et al. "Modulation of gut microbiota during probiotic-mediated attenuation of metabolic syndrome in high fat diet-fed mice". In: *The ISME Journal* 9.1 (Jan. 2015), pp. 1–15. DOI: 10.1038/ismej.2014.99.

[6] Steenbergen L et al. "A randomized controlled trial to test the effect of multispecies probiotics on cognitive reactivity to sad mood". In: *Brain, Behavior, and Immunity* 48 (Aug. 2015), pp. 258–264. DOI: 10.1016/j.bbi.2015.04.003.

[7] Cho I and Blaser MJ. "The human microbiome: at the interface of health and disease". In: *Nature Reviews Genetics* (Mar. 2012). DOI: 10.1038/nrg3182.

[8] Dodd D, Tropini C, and Sonnenburg JL. "Your gut microbiome, deconstructed". In: *Nature Biotechnology* 33.12 (Dec. 2015), pp. 1238–1240. DOI: 10.1038/nbt.3431.

[9] Nielsen HB et al. "Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes." In: *Nature biotechnology* 32.8 (2014), pp. 822–828. DOI: 10.1038/nbt.2939.

[10] Faith JJ et al. "Identifying Gut Microbe-Host Phenotype Relationships Using Combinatorial Communities in Gnotobiotic Mice". In: *Science Translational Medicine* 6.220 (Jan. 2014), 220ra11–220ra11. DOI: 10.1126/scitranslmed.3008051.

[11] Brestoff JR and Artis D. "Commensal bacteria at the interface of host metabolism and the immune system". In: *Nature Immunology* 14.7 (June 2013), pp. 676–684. DOI: 10.1038/ni.2640.

[12] Schaedler RW. "ASSOCIATION OF GERMFREE MICE WITH BACTERIA ISOLATED FROM NORMAL MICE". In: *Journal of Experimental Medicine* 122.1 (July 1965), pp. 77–82. DOI: 10.1084/jem.122.1.77.

[13] Dewhirst FE et al. "Phylogeny of the defined murine microbiota: altered Schaedler flora." In: *Applied and environmental microbiology* 65.8 (Aug. 1999), pp. 3287–92.

[14] Geuking MB et al. "Intestinal Bacterial Colonization Induces Mutualistic Regulatory T Cell Responses". In: *Immunity* 34.5 (May 2011), pp. 794–806. DOI: 10.1016/j.immuni.2011.03.021.

[15] Wymore Brand M et al. "The Altered Schaedler Flora: Continued Applications of a Defined Murine Microbial Community". In: *ILAR Journal* 56.2 (2015), pp. 169–178. DOI: 10.1093/ilar/ilv012.

[16] Sarma-Rupavtarm RB et al. "Spatial distribution and stability of the eight microbial species of the altered schaedler flora in the mouse gastrointestinal tract." In: *Applied and environmental microbiology* 70.5 (2004), pp. 2791–800. DOI: 10.1128/AEM.70.5.2791-2800.2004.

[17] Ge Z et al. "Colonization Dynamics of Altered Schaedler Flora Is Influenced by Gender, Aging, and Helicobacter hepaticus Infection in the Intestines of Swiss Webster Mice". In: *Applied and Environmental Microbiology* 72.7 (July 2006), pp. 5100–5103. DOI: 10.1128/AEM.01934-05.

[18] Stehr M et al. "Charles River altered Schaedler flora (CRASF(R)) remained stable for four years in a mouse colony housed in individually ventilated cages". In: *Laboratory Animals* 43.4 (Oct. 2009), pp. 362–370. DOI: 10.1258/la.2009.0080075.

[19] Singer SM and Nash TE. "The Role of Normal Flora in Giardia lamblia Infections in Mice". In: *The Journal of Infectious Diseases* 181.4 (Apr. 2000), pp. 1510–1512. DOI: 10.1086/315409.

[20] Shen TCD et al. "Engineering the gut microbiota to treat hyperammonemia". In: *Journal of Clinical Investigation* 125.7 (July 2015), pp. 2841–2850. DOI: 10.1172/JCI79214.

[21] Collins J et al. "Intestinal microbiota influence the early postnatal development of the enteric nervous system". In: *Neurogastroenterology & Motility* 26.1 (Jan. 2014), pp. 98–107. DOI: 10.1111/nmo.12236.

[22] Henderson AL et al. "Attenuation of Colitis by Serum-Derived Bovine Immunoglobulin/Protein Isolate in a Defined Microbiota Mouse Model". In: *Digestive Diseases and Sciences* 60.11 (Nov. 2015), pp. 3293–3303. DOI: 10.1007/s10620-015-3726-5.

[23] Moghadamrad S et al. "Attenuated portal hypertension in germ-free mice: Function of bacterial flora on the development of mesenteric lymphatic and blood vessels". In: *Hepatology* 61.5 (May 2015), pp. 1685–1695. DOI: 10.1002/hep.27698.

[24] Wannemuehler MJ, Overstreet AM, Ward DV, and Phillips GJ. "Draft Genome Sequences of the Altered Schaedler Flora, a Defined Bacterial Community from Gnotobiotic Mice". In: *Genome Announcements* 2.2 (2014), e00287–14–e00287–14. DOI: 10.1128/genomeA.00287-14.

[25] Wang J et al. "PNAS Plus: From the Cover: Dietary history contributes to enterotype-like clustering and functional metagenomic content in the intestinal microbiome of wild mice". In: *Proceedings of the National Academy of Sciences* 111.26 (July 2014), E2703–E2710. DOI: 10.1073/pnas.1402342111.

[26] Eddy SR. "Profile hidden Markov models". In: *Bioinformatics* 14.9 (Oct. 1998), pp. 755–763. DOI: 10.1093/bioinformatics/14.9.755.

[27] Powell S et al. "eggNOG v4.0: nested orthology inference across 3686 organisms". In: *Nucleic Acids Research* 42.D1 (2014), pp. D231–D239. DOI: 10.1093/nar/gkt1253.

[28] Nguyen TLA, Vieira-Silva S, Liston A, and Raes J. "How informative is the mouse for human gut microbiota research?" In: *Disease Models & Mechanisms* 8.1 (Jan. 2015), pp. 1–16. DOI: 10.1242/dmm.017400.

[29] Jensen PA, Dougherty BV, Moutinho TJ, and Papin JA. "Miniaturized Plate Readers for Low-Cost, High-Throughput Phenotypic Screening". In: *Journal of Laboratory Automation* 20.1 (2015), pp. 51–55. DOI: 10.1177/2211068214555414.

[30] Veselkov KA et al. "Recursive Segment-Wise Peak Alignment of Biological 1 H NMR Spectra for Improved Metabolic Biomarker Recovery". In: *Analytical Chemistry* 81.1 (Jan. 2009), pp. 56–66. DOI: 10.1021/ac8011544.

[31] Oberhardt MA et al. "Genome-scale metabolic network analysis of the opportunistic pathogen Pseudomonas aeruginosa PAO1." In: *Journal of bacteriology* 190.8 (Apr. 2008), pp. 2790–803. DOI: 10.1128/JB.01583-07.

[32] Overbeek R. "The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes". In: *Nucleic Acids Research* 33.17 (2005), pp. 5691–5702. DOI: 10.1093/nar/gki866.

[33] Robertson BR. "Mucispirillum schaedleri gen. nov., sp. nov., a spiral-shaped bacterium colonizing the mucus layer of the gastrointestinal tract of laboratory rodents". In: *INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY* 55.3 (May 2005), pp. 1199–1204. DOI: 10.1099/ijs.0.63472-0.

[34] Ghosh A et al. "A Peptide-Based Method for 13C Metabolic Flux Analysis in Microbial Communities". In: *PLoS Computational Biology* 10.9 (Sept. 2014). Ed. by Ouzounis CA, e1003827. DOI: 10.1371/journal.pcbi.1003827.

[35] Arquiza JA and Hunter J. "The use of real-time PCR to study Penicillium chrysogenum growth kinetics on solid food at different water activities". In: *International Journal of Food Microbiology* 187 (Sept. 2014), pp. 50–56. DOI: 10.1016/j.ijfoodmicro.2014.06.002.

[36] Novakova J et al. "Selective growth-inhibitory effect of 8-hydroxyquinoline towards Clostridium difficile and Bifidobacterium longum subsp. longum in co-culture analysed by flow cytometry". In: *Journal of Medical Microbiology* 63.Pt_12 (Dec. 2014), pp. 1663–1669. DOI: 10.1099/jmm.0.080796-0.

[37] Aghababaie M, Khanahmadi M, and Beheshti M. "Developing a kinetic model for co-culture of yogurt starter bacteria growth in pH controlled batch fermentation". In: *Journal of Food Engineering* 166 (Dec. 2015), pp. 72–79. DOI: 10.1016/j.jfoodeng.2015.05.013.

[38] Wolfe BE, Button JE, Santarelli M, and Dutton RJ. "Cheese Rind Communities Provide Tractable Systems for In Situ and In Vitro Studies of Microbial Diversity". In: *Cell* 158.2 (July 2014), pp. 422–433. DOI: 10.1016/j.cell.2014.05.041.

[39] Faust K and Raes J. "Microbial interactions: from networks to models". In: *Nature Reviews Microbiology* 10.8 (July 2012), pp. 538–550. DOI: 10.1038/nrmicro2832.

[40] Khare A and Tavazoie S. "Multifactorial Competition and Resistance in a Two-Species Bacterial System". In: *PLOS Genetics* 11.12 (Dec. 2015). Ed. by Zhang J, e1005715. DOI: 10.1371/journal.pgen.1005715.

[41] Lawrence D et al. "Species Interactions Alter Evolutionary Responses to a Novel Environment". In: *PLoS Biology* 10.5 (May 2012). Ed. by Ellner SP, e1001330. DOI: 10.1371/journal.pbio.1001330.

[42] Sibley CD et al. "Discerning the Complexity of Community Interactions Using a Drosophila Model of Polymicrobial Infections". In: *PLoS Pathogens* 4.10 (Oct. 2008). Ed. by Schneider DS, e1000184. DOI: 10.1371/journal.ppat.1000184.

[43] Klitgord N and Segrè D. "Environments that Induce Synthetic Microbial Ecosystems". In: *PLoS Computational Biology* 6.11 (Nov. 2010). Ed. by Papin JA, e1001002. DOI: 10.1371/journal.pcbi.1001002.

[44] Chiu HC, Levy R, and Borenstein E. "Emergent biosynthetic capacity in simple microbial communities." In: *PLoS computational biology* 10.7 (July 2014), e1003695. DOI: 10.1371/journal.pcbi.1003695.

[45] Traxler MF et al. "Interspecies interactions stimulate diversification of the Streptomyces coelicolor secreted metabolome". In: *mBio* 4.4 (2013). DOI: 10.1128/mBio.00459-13.

[46] Gefen O, Fridman O, Ronin I, and Balaban NQ. "Direct observation of single stationary-phase bacteria reveals a surprisingly long period of constant protein production activity". In: *Proceedings of the National Academy of Sciences* 111.1 (Jan. 2014), pp. 556–561. DOI: 10.1073/pnas.1314114111.

[47] Plata G, Henry CS, and Vitkup D. "Long-term phenotypic evolution of bacteria". In: *Nature* (2014). DOI: 10.1038/nature13827.

[48] Bartell JA et al. "Comparative Metabolic Systems Analysis of Pathogenic Burkholderia". In: *Journal of Bacteriology* 196.2 (2014), pp. 210–226. DOI: 10.1128/JB.00997-13.

[49] D'Souza G et al. "LESS IS MORE: SELECTIVE ADVANTAGES CAN EXPLAIN THE PREVALENT LOSS OF BIOSYNTHETIC GENES IN BACTERIA". In: *Evolution* 68.9 (Sept. 2014), pp. 2559–2570. DOI: 10.1111/evo.12468.

[50] Donohoe DR et al. "A Gnotobiotic Mouse Model Demonstrates That Dietary Fiber Protects against Colorectal Tumorigenesis in a Microbiota- and Butyrate-Dependent Manner". In: *Cancer Discovery* 4.12 (Dec. 2014), pp. 1387–1397. DOI: 10.1158/2159-8290.CD-14-0501.

[51] Smith PM et al. "The Microbial Metabolites, Short-Chain Fatty Acids, Regulate Colonic Treg Cell Homeostasis". In: *Science* 341.6145 (Aug. 2013), pp. 569–573. DOI: 10.1126/science.1241165.

Chapter 8

Reflections and Future Directions

Over the course of completing Aims 1–3, I developed novel multiscale models which integrate metabolic networks with other relevant data on microbial communities, I leveraged information inherent in metabolic networks to improve the assembly of microbial genomes from metagenomic data, and I demonstrated how ensembles of metabolic networks are an effective approach to managing uncertainty in metabolic network structure. In additional, related work, I spear-headed an effort to characterize the altered Schaedler flora (ASF) and further develop it as a model microbiota. Having completed this work, I now turn to describing how each project advanced my field in key ways. In this final section I attempt to articulate my perspective on the work reported in this dissertation, and I discuss the next steps for developing computational models of the microbiota into tools that will benefit society.

8.1 Evaluating impact and looking forward

Computational models of the gut microbiome in the clinic

The overarching motivation behind my work has been a long-term view of the future of computational models of the human microbiome, and the positive impact that those computational models can have on human well-being. Eventually, I expect that it will be possible to generate highly accurate physiological models for any microbial species based exclusively on the genome, and that it will be possible to accurately predict microbe-microbe and microbe-host interactions based on genomic information such as antimicrobial peptides produced or sensitivity to such peptides, the influence of quorum sensing and other signaling processes, the spatial distribution of the microbial community and host cells, the metabolic environment, and many other important factors. Based on these sources of information, then it will be possible, in principle, to use a computational model of the gut microbiome to rationally design improved therapeutic strategies and pharmaceuticals. Such models will be useful for drug discovery by identifying highly-relevant drug targets. Such models will be useful for regulatory approval of therapeutic approaches or pharmaceuticals (or at least as a pre-clinical trial step) in much the same way that a computational model of the glucose-insulin system is used by the FDA to test closed-lop insulin pumps [1]. Perhaps such models will also be used in a personalized medicine context by integrating patient 'omics data and suggesting the optimal drug, pre- or probiotic treatments. Eventually, mechanistic computational models based on metabolic networks will be powerful clinical tools.

While this long view is the motivation for my work—to develop tools which will allow genome-scale metabolic networks to impact the clinic in the ways just described—there are substantial obstacles to arriving at that point (e.g. present difficulties in annotating genes, the stochastic element in microbiome dynamics, the many non-metabolic factors which influence microbe-microbe interactions and microbe-host interactions). In the short-term, genome-scale metabolic networks are exeptionally valuable as research tools, and my work provides stepping stones to the exciting future described above.

Mechanistic models are powerful research tools

In a research setting, genome-scale metabolic networks serve two primary roles. First, they serve as a formalized repository of all knowledge about biochemical conversions that can be performed by a given organism. Second, that knowledge is represented in a way amenable to simulation, which is a formalized way to generate hypotheses about an organism's metabolic function. Viewed in this way, the value of mechanistic models is not exlusively in the ability to make accurate predictions, but additionally, in the ability to represent knowledge in a useful way and to identify the natural conclusions that arise from that knowledge as it accumulates. A prediction from a metabolic network can be correct (supported by experimental data) or not (disagrees with experimental data), but either outcome is valuable to the research process. For example, an incorrect prediction indicates an incomplete or incorrect understanding of the system. New simulations can be performed to identify potential fixes to reconcile predictions and experimental data. Those potential fixes then serve as hypotheses which can be tested, which leads to new knowledge.

The value and future of mutliscale models of microbial communities

In a research setting, multiscale models serve the same primary roles as genome-scale metabolic networks. Really, any mechanistic computational model of a poorly-characterized biological system is most valuable in these two roles (by mechanistic, I mean the elements of the model correspond directly to elements in the system). Multiscale models serve as a resource for representing knowledge of how a systems works at different scales, and for generating testable hypotheses that arise from the current understanding. 'Omics data sets are accumulating as our ability to characterize microbial communities at every scale increases, and we know that every 'omics perspective is dependent on others (e.g. proteomics measures protein abundances, which are dependent on mRNA transcripts, measurable by transcriptomics). Additionally, macro-scale measurements of microbial communities (such as species growth rates) are highly dependent on the lower-level biological processes. Multiscale models will become increasingly essential for integrating these large-scale, complext data sets and macro-scale measurements, and generating useful hypotheses from them. Additionally, I expect that as 'omics data becomes easier to to generate, time-series data sets will be come more common, and the need for dynamic multiscale models (as presented in Chapter 1) will grow.

The value of my work in Aim 1 is in the development of new ways to represent knowledge at multiple scales and to generate hypotheses from that knowledge.

In Chapter 2 I presented a multiscale model of *P. aeruginosa* biofilm formation which integrated knowledge about *P. aeruginosa* metabolism with knowledge about the biofilm assembly rules from the standpoint of the individual cells. To my knowledge, that was the first time genome-scale metabolic networks had been integrated with information about the spatial distribution of the community of cells. The following year, a similar modeling framework was developed to simulate interactions between bacterial colonies on an agar plate [2]. Without considering the resulting predictions, the modeling framework itself is a conceptual advance that can contribute to the study of any microbial community where both chemical and spatial factors influence community development over time.

An obstacle to the application of this particular multiscale modeling framework is the computational complexity required to run simulations. The need to solve repetitive linear programs (simulating metabolic processes) and numerically simulate diffusion of more than 100 metabolites required 15+ hours to run every individual simulation. This computational burden can be prohibitive to performing enough replicates of a simulation to accurately estimate variance. Another concern with this and other multiscale modeling frameworks is that the model can become overly complex. The abundance of parameters can lead to overfitting available data, or to such complex simulations that interpreting the results can be just as challenging as interpreting the results of an actual experiment in the lab. The answer to these concerns is that a model should only be as complicated as needed to address a particular problem, and that appropriately complex multiscale models can still lead to biological discovery [2]. Future research could beneficially be aimed at increasing the efficiency of various elements of the multiscale model, including dynamic flux balance analysis and diffusion simulations.

Moving forward, I expect that some form of this model will be important to studying the dynamics of gut microbial communities. In the gut, unique microbes live higher up in the GI tract with unique adaptations to extremes in pH, and there are distinct changes in community composition moving from small intestine to colon [3]. There is additional structure starting from the endothelial cells of the intestinal wall up through the mucus layer into the intestinal contents [4]. Metabolites and cells are carried along by peristaltic forces, and diffusion carries metabolites to and from the intestinal lining. The interplay of all these spatial factors creates many unique and dynamic micro-environments, which in turn interact with the metabolic process of each unique species. A multiscale model of this system would open the possibility of exploring unique questions. For example, what factors in the system have the strongest impact on host metabolism? A systematic sensitivity analysis could identify the microbial taxa, spatial regions, and spatial parameters (such as the velocity of lumenal contents) which most impact the flow of nutrients to the host. To build a large-scale model like this, it would be prudent to start with a simpler system such as an artificial gut reactor [5] or the gut microbiome of gnotobiotic mice [6]. The initial goal would be to maintain the model as a knowledge repository and a method of generating hypotheses. Such a pairing between a multiscale model and an experimental system would serve as an additional stepping stone towards powerful clinical tools.

In Chapter 3 I presented a microbial interaction network inferred from time series metagenomic data. In that work I integrated predictions from genome-scale metabolic networks with the inferred interactions of the network in order to hypothesize about mechanisms of interaction. It was not a multiscale model in the same sense as the biofilm model earlier, but it was a new way

to integrate data from different scales together and make hypotheses based on the available information. In terms of the two primary roles I mentioned above, the work in Chapter 3 addresses the second (i.e. a need for new ways to generate hypotheses). The role of metabolic networks in Chapter 3 was to estimate the role of metabolism in explaining the inferred interactions. I chose to represent genus-level metabolism as the union of many species-level metabolic networks. This was intended to represent the scope of the genus "pan-genome", but an alternative approach would be to represent a genus-level metabolic network as an ensemble (a different application of ensembles than discussed in Chapter 6). Interactions between genera could then be represented as a distribution of interactions between the species-level networks in each genus-level ensemble. This is a potentially fruitful future direction.

The value and future of SONEC

In Chapter 5 I presented SOrting by NEtwork COmpletion (SONEC), an approach for binning short metagenomic sequence fragments into species-specific genomes. This is a fundamentally different role for metabolic networks than has been demonstrated previously. The SONEC algorithm was intended to fill a need arising from Aim 1: that is, in order to build multiscale models of real microbiomes, we first need genomes for each species present. As I mentioned in the context for that chapter, there are now sequencing technologies available which significantly reduce the need for metagenomic read binning approaches such as SONEC, including single-cell [7] and long-read sequencing technologies [8]. Therefore, the value of SONEC now is not so much in the specific algorithm itself (which may find limited use in traditional, short-read metagenomic data sets), but rather in the concept that biochemical networks can serve as an additional source of information for finding order in chaotic meta-omics data. Other applications where a similar concept could be applicable are metabolomics and proteomics. I can imagine a case where draft metabolic networks could be reconstructed for species in a community based on metagenomic data, and those models would be used to help interpret additional meta-omics data, such as identifying the species most likely to have produced a particular metabolite or peptide.

The value and future of Ensemble FBA

In Chapter 6, I introduced Ensemble FBA, which allows for the representation and analysis of many competing hypotheses with regard to metabolic network structure. I believe that ensemble analysis of metabolic networks has improved—and with additional research, will continue to improve—the ability to predict microbial physiological properties based on limited information (a key challenge in computational biology). One possible direction for future work would be to develop optimal methods for deciding which networks within the ensemble are closest to the true network structure. For example, I explored the possibility of designing an alorithm which would choose an optimal sequence of experiments to systematically narrow down the network structures. In this way, an ensemble could lead to more efficient use of wet lab resources. A future direction that would further apply ensembles to managing uncertainty in network structure would be to integrate regulatory networks (or ensembles of regulatory networks) with ensembles of metabolic networks. This would be one way to represent the space of possible network structure/state combinations given available information. Finally, an important future direction will be to explore alternative methods to generate ensembles of metabolic networks, including sampling in probability-based reconstruction methods, or generating individual networks using different reconstruction methods altogether.

Another important result from the Ensemble FBA project was the trimming algorithm, devised to make better use of carbon utilization data. We found that by incorporating negative growth conditions using this new trimming algorithm, ensemble accuracy was improved by $\sim$15%. However, I did not take into account the annotation confidence when trimming reactions, and possibly trimmed reactions which are actually active in the organism. One possible direction for future work is to weight the reactions by how confident the gene annotation was, so that low-confidence reactions are more likely to be removed. Another possible direction is to incorporate the trimming algorithm with regulatory information, to identify genes whose expression is most likely suppressed under particular growth conditions.

Ensemble FBA can improve the predictions generated by metabolic network analysis. Improved predictions benefit the integration of metabolic networks into multiscale models. Future efforts can be directed towards improving ensemble-level predictions, including weighting network votes or treating the distribution of votes probabilistically.

The value and future of model microbiota paired with computational models

In Chapter 7 I presented additional work characterizing the altered Schaedler flora (ASF). The intention is that the ASF serve as a platform for validating metabolic network-based models of the gut microbiome. The gut

microbiome is still a poorly-characterized system, even given a mere 8 ASF species in a gnotobiotic mouse. By first characterizing the ASF gut microbiome, we expect that important principles will be discovered which are applicable to more complex gut microbiomes. Genome-scale metabolic networks will serve as a repository of knowledge as our efforts to characterize the ASF advance, and analysis of the network models will guide experimentation. Some questions we hope to address would include: What is the role of each species in the community in terms of host nutrition and community structure? A related question, what happens when each species is removed (in terms of community structure and metabolic function)? What happens when the diet is changed, and how are the changes a function of the metabolism of each species? How do nutrients flow through the system, and how much of the bioconversions we observe can be predicted by metabolic models?

As the ASF microbiome becomes better characterized (which we expect to happen much faster with the ASF than with a more complex microbiome), we can move towards testing new algorithms that could someday have clinical value. The ASF will serve as a sandbox for testing new pre- and probiotic design algorithms. For example, an arbitrary objective could be to design prebiotics which increase the abundance of *Parabacteroides* ASF519 in the small intestine. Starting with several candidate pre- or probiotics suggested by one or more algorithms, it would be much more straightforward to test and evaluate those treatments in ASF gnotobiotic mice than in other, more complex systems. Even with a simple community such as the ASF, I would expect this process to be difficult. As mentioned at the beginning of this chapter, there are many non-metabolic factors that determine microbial interactions. Even predicting metabolic interactions is still difficult to do without prior knowledge about the community [9]. A large part of working with a simple community such as the ASF is being able to identify those factors , metabolic or not (e.g. host diet, species abundances, non-metabolic interactions), which are most predictive of desired outcomes (e.g. metabolite levels in host blood, host weight, abundances of particular community members) and devising ways to incorporate those important factors into the modeling framework.

The results from Aims 1–3 fit into a larger workflow

Conceptually, Aim 2 (SONEC) improved the first step in a genome-scale metabolic network analysis, which is to gain access to a genome. Aim 3 (EnsembleFBA) improves the predictions that can be made from genomic information. Both Aim 2 and Aim 3 support Aim 1 (multiscale modeling) by providing genomes and improving the models derived from those genomes, which then improves the predictions from any derivative multiscale models.

8.2 Conclusion

The human microbiome exerts an enormous influence on host health. Genome-scale metabolic networks are powerful research tools for tracking knowledge and systematically generating testable hypotheses. My graduate work has resulted in advances in the ability to incorporate metabolic networks into rich multiscale models. My work has also resulted in the SONEC algorithm, a demonstration of how metabolic networks can help to interpret meta-omics data. Ensemble FBA is a key result that improves the predictions that can be generated from draft metabolic networks. By beginning to characterize the ASF, I have furthur molded the ASF into a useful experimental tool for validating computational methods and discovering important principles of gut microbiome function. In summary, I have introduced conceptual advances and concrete tools which will serve as stepping stones towards the larger goal of developing robust computational biology tools for engineering microbial communities of societal importance.

8.3 References

[1] Winslow RL, Trayanova N, Geman D, and Miller MI. "Computational Medicine: Translating Models to Clinical Care". In: *Science Translational Medicine* 4.158 (Oct. 2012), 158rv11–158rv11. DOI: 10.1126/scitranslmed.3003528.

[2] Harcombe WR et al. "Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics." In: *Cell reports* 7.4 (May 2014), pp. 1104–15. DOI: 10.1016/j.celrep.2014.03.070.

[3] Sarma-Rupavtarm RB et al. "Spatial distribution and stability of the eight microbial species of the altered schaedler flora in the mouse gastrointestinal tract." In: *Applied and environmental microbiology* 70.5 (2004), pp. 2791–800. DOI: 10.1128/AEM.70.5.2791-2800.2004.

[4] Zoetendal EG et al. "Mucosa-Associated Bacteria in the Human Gastrointestinal Tract Are Uniformly Distributed along the Colon and Differ from the Community Recovered from Feces". In: *Applied and Environmental Microbiology* 68.7 (July 2002), pp. 3401–3407. DOI: 10.1128/AEM.68.7.3401-3407.2002.

[5] McDonald JA et al. “Simulating distal gut mucosal and luminal communities using packed-column biofilm reactors and an in vitro chemostat model”. In: *Journal of Microbiological Methods* 108 (Jan. 2015), pp. 36–44. DOI: 10.1016/j.mimet.2014.11.007.

[6] Wymore Brand M et al. “The Altered Schaedler Flora: Continued Applications of a Defined Murine Microbial Community”. In: *ILAR Journal* 56.2 (2015), pp. 169–178. DOI: 10.1093/ilar/ilv012.

[7] Yilmaz S and Singh AK. “Single cell genome sequencing”. In: *Current Opinion in Biotechnology* 23.3 (June 2012), pp. 437–443. DOI: 10.1016/j.copbio.2011.11.018.

[8] Koren S and Phillippy AM. “One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly”. In: *Current Opinion in Microbiology* 23 (Feb. 2015), pp. 110–120. DOI: 10.1016/j.mib.2014.11.014.

[9] Zomorrodi AR and Maranas CD. “OptCom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities.” In: *PLoS computational biology* 8.2 (Feb. 2012), e1002363. DOI: 10.1371/journal.pcbi.1002363.