Modeling Host-Microbiome Interactions

BY

PETER E LARSEN
B.S., Purdue University 1993
M.S., University of Illinois at Chicago 2006

THESIS

Submitted as partial fulfilment of the requirements
for the degree of Doctor of Philosophy in Bioinformatics
in the Graduate College of the University of Illinois at Chicago, 2017

Chicago, Illinois

Defense Committee:

Dr. Yang Dai:	Chair and Advisor
Dr. Giamila Fantuzzi:	Kinesiology and Nutrition
Dr. Rachel Poretsky:	Biological Sciences
Dr. Ao Ma:	Bioengineering
Dr. Dionysios Antonopoulos:	Dept. Medicine, University of Chicago

This thesis is dedicated to my wife, Danielle.

ACKNOWLEDGEMENTS

I would like to thank my advisor and collaborator, Dr. Yang Dai, who has shown exceptional patience in helping me translate my thoughts into rigorous computational objectives.

I would like to thank my committee members Dr. Giamila Fantuzzi, Dr. Rachel Poretsky, Dr. Ao, and Dr. Dionysios Antonopoulos for much valuable input.

I would also like to acknowledge Dr. Lawrence David at the Duke Center for Genomics and Computational Biology for generously making his human microbiome dataset available to me.

TABLE OF CONTENTS

CHAPTER PAGE

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF ABBREVIATIONS

16S rRNA	30S small subunit of prokaryotic ribosome
BL6	Mouse strain C57ML/6J
CIN	Community Interaction Network
COX	Cyclooxygenase
DBN	Dynamic Bayesian Network
EFP	Enzyme Function Profile
FBA	Flux Balance Analysis
GA	Genetic Algorithm
HF	High Fat
HFS	High Fat diet Score
LF	Low Fat
LOO-CV	Leave One Out Cross Validation
MAP	Microbial Assemblage Prediction
MCC	Mathew's Correlation Coefficient
MI	Machine Intelligence
OS	Obesity Score
OTU	Operational Taxonomic Unit
PCC	Pearson's Correlation Coefficient
PRMT	Predicted Relative Metabolic Turnover
PRTT	Predicted Relative Transmembrane Transport
RHF	Relative expectation of High Fat diet
RPO	Relative Propensity for Obesity
Sv129	Mouse strain 129S6/AvEvTac
SVM	Support Vector Machine
TAP	Taxonomic Average function Profile
TAvP	Taxonomic Average and Variance function Profile

SUMMARY

One of the most important discoveries that has come from recent advances in DNA sequencing technology is finding bacteria present in complex communities almost everywhere we have looked for it. Notable for its relevance to human health, one of the places that has been found to harbor rich ecosystems of bacteria is the human body. In fact, every human individual has from three to five pounds of bacteria living on and inside them. We term these populations of bacteria the 'microbiome'. These populations of bacteria are not all harmful to human health and many are in fact beneficial, if not crucial, to their human host. These bacteria help us to fight off infections, assist us in digesting our food, and forge close associations with our healthy immune systems. When these communities are out of balance however, they can also contribute to making us sick, predisposing us to obesity, increasing our risk of cancer, and leading to inflammatory diseases such as Irritable Bowel Syndrome. Understanding how these recently discovered communities of bacteria interact with you is important to your health and may provide tools for fighting disease. Analyzing these bacterial communities however can be very difficult. A single teaspoon of a microbiome might include over ten billion bacteria, and few if any of these bacteria can be grown and studied in a biological laboratory. Studying these bacterial communities requires the application of advanced DNA sequencing technologies and complex computer analysis approaches.

In this study "Modeling Host-Microbiome Interactions", we present a series of computational analyses that span multiple biological scales. In the first analysis, we computationally explore the interaction of a single group of bacteria, the Pseudomonads, with both human and plant hosts. In the second analysis, we examine the relationships of two human volunteers with their bacterial communities, how those communities can change in response to what their host eats, and how those communities can make their hosts sick. In the third analysis, we use what we learned in the first two analyses to investigate a series of larger experiments using laboratory mice to investigate the interactions between a mouse and its microbiome. Using the mouse microbiome data, we build predictive models of how a microbiome changes

SUMMARY (CONTINUED)

with the mouse’s diet, models to predict what sort of metabolic activities might be occurring within a mouse’s microbiome, and a computational model that predict the effects of a mouse’s microbiome on the likelihood that the mouse will be predisposed to obesity. When we assemble these models into a single computational tool, a computational model we call the iMOUSE, we can perform ‘experiments’ on this simulated mouse that closely match the results that are observed in real-world laboratory mouse experiments. Using the iMOUSE model, we can perform many millions of simulated experiments very quickly, allowing us to investigate the interactions between mice and their microbiome. We have used this computational model to search many millions of possible diets to find a diet that will reduce the risk of obesity in a mouse with a microbiome that predisposed it to gain weight.

The success of our iMOUSE model provides a potentially powerful tool for investigating the interactions between a host and its microbiome. This tool will help identify methods for improving human health and fighting disease through understanding and manipulating your symbiotic bacterial communities.

1. INTRODUCTION

We live in a world dominated by the action of bacterial communities. By biomass, bacteria are the planet earth's dominant form of life (Whitman et al. 1998). Not only do microbial communities influence all of the planet's biogeochemical cycles (O'dor et al. 2009, Hoegh-Guldberg 2010), but bacterial communities also form tightly knit symbiotic interactions with more complex organisms (Bourne et al. 2008, Ezenwa et al. 2012, Wylie et al. 2012, Yoon et al. 2015). While these symbiotic communities have been with us and around us all along, they have until recently remained nearly opaque to methodical investigation. The vast majority of bacteria cannot be easily cultured and characterized in the laboratory and cannot be well described by microscopy techniques. It is only through the advent of ultra-high throughput DNA sequencing of community-derived DNA that these communities have now become amenable to investigation (Shannon et al. 2003, Gilbert and Dupont 2011, Wang et al. 2015). These communities can be described in a number of ways: The community structure is the type and relative abundance of bacterial species present in a community; the metagenome is the type and abundance of specific genes or gene functions present in an entire population; the proteome is the abundance and predicted function of all of the proteins being expressed in a microbiome community; the community metabolome is the total of all the metabolic and transport reactions that can occur in a population, as encoded in the metagenome.

Bacterial communities have been found in nearly every environmental niche, from boiling water (Hugenholtz et al. 1998, Barns et al. 1999) to within millennia-old ice (Shtarkman et al. 2013), from surviving deep in the earth's crust (Takai et al. 2001, Edwards et al. 2006, Teske and Sorensen 2008) to living among the clouds (Fierer et al. 2008, Bowers et al. 2009). One important collection of environments that have been discovered to be home to a rich and varied microbial ecology is the human body and human associated environments (Wylie et al. 2012, Lax et al. 2014, Lax et al. 2017). Humans exist, we have discovered, not as individuals, but as superorganisms comprised of human cells that live in an inseparable symbiotic relationship with a vast ecosystem of microorganisms. By some measures, humans are actually in the minority in this superorganism. The number of human cells is outnumbered by bacterial cells in this ecosystem by more than an order of magnitude. The number of bacterial genes in this ecosystem outnumbers human genes by several orders of magnitude (Ding and Schloss 2014). These human-associated communities are collectively referred to as the human microbiome. Large-scale studies for identifying and characterizing microbiome communities, such as the Metagenomics of the Human Intestinal Tract (MetaHIT) project (http://www.metahit.eu/), the American Gut Project (http://americangut.org/), and the Human Microbiome Project (HMP) (http://www.hmpdacc.org/), have each contributed to our understanding of the relationships between microbiome community composition and the host. These studies have highlighted the microbiome as a valuable target for contributing to human health (Stulberg et al. 2016). These investigations in the microbiome have also highlighted that the tremendous diversity of the microbiome presents a significant challenge for analysis of human microbiome data.

1.1 Statement of the Problem: Understanding Disease in a Post-Koch's Postulates World

In 1884, Robert Koch and Friedrich Loeffler identified a set of criteria for linking a specific microorganism to a specific disease (Gradmann 2014). Koch's Postulates are:

1. The microorganism must be found in abundance in all organisms suffering from the disease, but should not be found in healthy organisms.
2. The microorganism must be isolated from a diseased organism and grown in pure culture.
3. The cultured microorganism should cause disease when introduced into a healthy organism.

4. The microorganism must be re-isolated from the inoculated, diseased experimental host and identified as being identical to the original specific causative agent.

Many of these postulates have been subject to significant modifications over time. For example, not all microbial infections lead to disease in all hosts and not all microorganisms, such as viruses, can be cultured. Nonetheless, these postulates, with modifications, have served as the foundation for microbiology for well over a century. As we develop the technology to explore the microbiome, however, we have found ourselves in a world where Koch's Postulates no longer strictly hold sway. We now know that many diseases are not caused by the presence of any single bacterium. Rather, disease can be defined as an emergent property of the microbiome community. An emergent property is a *categorial novum* that occurs when the properties of a system arise from interactions in the system's component elements, where no individual component of the system exhibits those properties. The whole is not only more, but in some way fundamentally different than the sum of its parts.

The question before us then is: How do we functionally and computationally define disease as an emergent property of the microbiome in the post-Koch's Postulate era and how do we use that understanding to propose methods to treat and prevent microbiome-based dysbiosis?

1.2 Purpose of the Study: Generating System-Scale Models of Host-Microbiome Interactions

The complex nature of host-microbiome interactions provides a multitude of specific opportunities to generate novel computational methods that elucidate how host health is influenced by the microbiome. A number of previously published host-microbe interaction experiments provide the necessary datasets for the development and validation of computational analysis methods. Each analysis in this research addresses host-microbiome interactions at a particular scale and, when taken together, comprises a coherent and novel framework for the identification of the molecular mechanisms of host-microbiome interactions and for hypothesizing methods of manipulating microbiome community structures to optimize host health.

The overarching goal of this research is to better understand, predict, and manipulate host-microbiome interactions to treat human disease and to promote health. The research has been divided into three aims, each aim spanning a specific scale of host-microbiome interactions. Together, these completed goal can be synthesized into a coherent framework to modeling and understanding the interactions between host microbiome and host phenotype as an emergent property of the microbiome community.

1.3 Significance of the Problem: Providing a Koch's Postulates for the Microbiome Era

The microbiome has been linked to a wide variety of diseases in humans. Associating a particular microbiome community structure to a specific disease state however is challenging. It has been reported that relevant patterns can be found among the highly varied microbiome communities. For example, a study of the microbiomes of a cohort of 4,788 microbiomes taken from 242 adults revealed that although community structures varied, specific metabolic pathways were found across multiple microbiome metagenomes (Huttenhower et al. 2012). In another study, it was reported that although the microbiome community structures of individuals and various sampled regions were distinct from one another, the community structures from one part of the body of an individual were predictive of the community structure of other body regions on the same individual (Ding and Schloss 2014). An individual's microbiome community structure is also dependent on the environment and the people, animals, and surfaces with which they interact (Lax et al. 2014, Lax et al. 2017).

Observing that there is a significant correlation between the relative abundances of some microbiome community members and human health does not necessarily identify the underlying molecular mechanisms driving this relationship. To leverage the microbiome community for the benefit of human health, any new analysis approaches will have to explore more than just the community structures of microbiomes to find biologically relevant patterns. It is necessary to generate predictive models, describing how microbiome communities interact with each other and their host to influence health.

The computational models outlined here are predicated on a novel insight very different from Koch's Postulates: The key mechanisms of host-microbiome interactions are not dependent upon the presence or absence of a single bacterial species or microbial genes. Rather, the effects of host microbiome interactions are an emergent property of the microbial community. Clearly defining analytical and computational approaches that link microbiomes to host health will provide the 'Koch's Postulates' for the 21st century and, perhaps, the next century of research and discovery.

1.4 Significance of the Study: New Tools for Analyzing the Emergent Properties of Microbiome Communities

In a recent National Science and Technology Council Committee of US government scientists (Stulberg et al. 2016), a "priority need for new tools, technologies, and databases" was identified as the most pressing need for advancing microbiome research. This research effort directly addresses the need for advanced computational tools by developing a set of novel, predictive microbiome analysis methodologies for omics microbiome data. The analysis tools can be further utilized to generate system-scale models of host-microbiome interactions and directly propose additional biological experiments and, potentially novel therapeutic interventions for manipulating the microbiome to drive a desired host phenotype.

In a recent review of microbiome and metabolic diseases, specific genera, classes, or species of bacteria cannot be positively or negatively correlated with specific host phenotypes (Fukuda and Ohno 2014). This suggests that to confidently link microbiomes with host-microbiome interactions, more information than microbiome community structure is needed. This need is directly addressed in the system-scale computational tools developed here, surpassing the previous limitations of microbiome analysis.

In addition to the significance of system-scale models for host-microbiome interactions to scientific and clinical milieus, this dissertation also extends several technical innovations including developing innovative approaches to modeling microbiome community dynamics, innovative statistical analysis approaches for predicting microbiome community enzyme function profiles from community structure, and a unique approach to quantifying a bacterial transcriptome from genomic sequence analysis.

1.5 Outline of Dissertation

The remainder of the dissertation is divided into six chapters.

Chapter 2 provides a brief review of available literature to provide context for host-microbiome interactions, methods for investigating the microbiome, and previous microbiome community modeling approaches.

A collection of computational analysis tools for investigating the microbiome that we have developed in previous research efforts and are either used or improved in our current research are described in **Chapter 3**.

In Chapters 4, 5, and 6, our computational models are presented, spanning multiple and expanding scales of host-microbiome interactions.

Chapter 4 is “Pseudomonas-Host Interactions”. Current high throughput sequencing technology can assemble genomes directly from environmentally collected DNA from otherwise uncharacterized and uncultured bacteria. In the case of human microbiome metagenomes, those uncharacterized bacteria might include antibiotic resistant pathogens or other hospital acquired infections. A method is needed for predicting a potentially newly discovered bacteria's ecological host-interaction niche from an annotated genome alone so that genomic information can be translated into effective therapeutic interventions.

We have chosen to focus on the bacteria Pseudomonads for our methods development and validations. Pseudomonads are ubiquitous components of environmental ecosystems as well as the second most commonly acquired opportunistic infection in hospitals (de Bentzmann and Plesiat 2011, Silby et al. 2011). From annotated genomes, metabolomic and transportomic models can be generated for Pseudomonads. We have developed and validated a novel transportomic modeling approach, namely, Predicted Relative Transmembrane Transport, for use in this analysis. Genomic, metabolomic, and transportomic modeling data can be analyzed using Support Vector Machines (SVMs) not only to predict if a sequenced Pseudomonad is likely to be a pathogen, but also to propose the molecular mechanisms of host-microorganism interaction. These predictions may lead to the identification of specific molecular targets for the development of new antibiotics and innovative therapies to treat antibiotic resistant hospital acquired infections.

Much of this analysis was initially published in (Larsen et al. 2014, 2015a) and uses many of our previously published tools in (Larsen et al. 2011).

Chapter 5 is "Modeling Human Dysbiosis". "Dysbiosis" is any condition in which perturbations to a host's microbiome leads to a negative impact on the host's health. Microbiome community structures, however, can differ vastly from host to host and over time within the same host, making it difficult to generalize the molecular mechanisms of host-microbiome interaction that lead to dysbiosis. We propose that it is not the microbiome community structure that is most informative of host-microbiome interaction, but rather the emergent properties of the microbiome community that must be considered. The emergent properties of the microbiome community structure are community enzyme function profiles, and community metabolome. SVMs can be used to predict host dysbiosis using community structure, enzyme function profile, or community metabolome. This method also proposes the specific organisms, enzyme activities, and metabolites that are predictive of host dysbiosis and hypothesizes sets of specific molecular targets for possible therapeutic interventions. When the SVM dysbiosis predictor is combined with a dynamic model of the effects of the host's diet on a microbiome community structure, a system-scale model of host-microbiome interactions is created with the capacity to accurately predict diet conditions that are likely to result in diet-induced dysbiosis.

Portions of this have been previously published in (Larsen and Dai 2015) and utilizes our previously published approaches in (Larsen et al. 2012a, Larsen et al. 2015b).

Chapter 6 is "Generating a System-Scale Model of a Mouse Obesity Host-Microbiome Interaction". The relationships between host, microbiome, and host phenotype are complex. While the opportunity to conduct longitudinal studies or hypothesis-driven experiments of human subjects is limited, the use of gnotobiotic mice, mice that have completely characterized microbiome compositions, provide a powerful experimental model for studying the interactions between host, microbiome, and the environment. The goal of this chapter is to develop a predictive model of host-microbiome interactions as a function of starting microbiome community structure and host diet, incorporating the metabolomic models and phenotype-predictions tools developed in Chapter 5. The result is a system-scale model of mouse-obesity host-microbiome interactions, called iMOUSE, which accurately predicts a series of previously published biological observations through *in silico* modeling. Of greatest value, the system model is capable of recapitulation of a set of experiments in which gnotobiotic mice with microbiomes transplanted from obese humans are themselves prone to obesity. This system-scale model can then be used to propose a specific diet that will manipulate the host-microbiome interactions to yield a non-obese phenotype for the host for any starting microbiome community structure.

Chapter 7 discusses the conclusions that can be drawn from this research, the potential weaknesses of the individual approaches, and outlines the future directions that the computational tools can take us.

2. BACKGROUND AND LITERATURE SURVEY

We become hosts to a vast community of microorganisms within the first few moments of our lives (Gritz and Bhandari 2015, Shin et al. 2015, Yang et al. 2016b), and these communities begin their work of our decomposition at the instant of our deaths (Metcalf et al. 2016). Between times, our microbiomes grow and change with us, molding the development of our immune systems, guiding our metabolisms, defending us from infection, and helping us to digest our food (Fukuda and Ohno 2014, Putignani et al. 2014, Yang and Jobin 2014, Aw and Fukuda 2015). It has often been written that the number of bacterial cells in the human body outnumber the human cells by as much as two orders of magnitude (Sender et al. 2016). Perhaps it would be more accurate to state that for every one hundred or so cells in a human, one or two cells are non-bacterial. In many ways, our microbiomes comprised of bacteria derived from a multitude of sources are as much a part of us, and what makes us human, as the eukaryotic cells we inherited from our parents.

Following, we discuss the diversity of the human microbiome, the ways that our microbiomes influence our health, and the mechanisms by which we can investigate the communities that make up our microbiomes.

2.1 Diversity of the Human Microbiome

Large-scale studies for identifying and characterizing microbiome communities such as the Metagenomics of the Human Intestinal Tract (MetaHIT) (http://www.metahit.eu/), the American Gut Project (http://americangut.org/), and the Human Microbiome Project (HMP) (http://hmpdacc.org/) have contributed to our understanding of the relationships between microbiome community composition and host. It has also highlighted that the tremendous diversity of microbiome populations presents a significant challenge in the analysis of human microbiome data. There are four phyla that predominate in the gut microbiome of most mammals, *Firmicutes*, *Bacteroides*, *Actinobacteria*, and *Proteobacteria* (Qin et al. 2010), but the actual number of species present can be in the thousands. A human microbiome is highly dynamic, changing in response to host behavior, environment, and diet (Theriot et al. 2014, Ursell et al. 2014, Rojo et al. 2015). Human microbiomes are so highly divergent from host to host, that it has been proposed that an individual might have a microbiome community structure that is unique (Kostic et al. 2013, Kostic et al. 2014). Host environment, diet, and genetics have been implicated in driving this diversity, although much of the variation among human microbiomes remains unexplained. The dynamic nature of these communities impedes our ability to make generalizations applicable across individual hosts and across microbiomes.

While a number of human 'ecosystems' have been well characterized, including skin, lung, and oral microbiomes (Huttenhower et al. 2012, Wylie et al. 2012, Aagaard et al. 2013, Ding and Schloss 2014), the largest component of our microbiome, by biomass, is the gut microbiome. The gut microbiome comprises 99% of the total microbial mass of the human microbiome (Schwabe and Jobin 2013). The GI tract is also one of the largest interfaces between the internal human environment and the rest of the world (Furness et al. 1999), making the gut microbiome one of the most important of out personal bacterial ecosystems. Additional information on the impact of the microbiome on their hosts can be found in Chapters 5 and 6.

2.2 The Microbiome Plays an Important Role in Human Health

The human microbiome provides many benefits to their hosts, including providing essential vitamins and nutrients and aiding in digesting proteins and complex carbohydrates (Bou Saab et al. 2014, Walsh et al. 2014), maintaining a healthy immune system (Calder 2013, Greer et al. 2013, Cantorna et al.

2014, Romano-Keeler and Weitkamp 2015) and defending against colonization by harmful or pathogenic organisms (Dutton and Turnbaugh 2012, Fuller 2012, Ramakrishna 2013, Hennessy et al. 2014), However, Dysbiosis occurs when perturbations in these microbiomes communities have a negative effect on the host's health (Tamboli et al. 2004, Yang and Jobin 2014, McLean et al. 2015). The negative effects that the microbiome can have on the host include increased propensity for obesity (Moran and Shanahan 2014, Sanz and Moya-Perez 2014, Cox and Blaser 2015), such as Irritable Bowel Syndrome (IBS) (Collins 2014, Dupont 2014, Kostic et al. 2014, Cammarota et al. 2015), increased vulnerability to cancers (Kipanyula et al. 2013, Schwabe and Jobin 2013, Viaud et al. 2014), and autoimmune disorders (Collado et al. 2015, McLean et al. 2015). Dysbiosis of the gut microbiome has been shown to coincide with the human host's mental health (Foster and McVey Neufeld 2013, Borre et al. 2014, Fond et al. 2015), including an increased risk of depression (Luna and Foster 2015).

Understanding the relationships between human health and the associated microbiome provides new and valuable tools for diagnostics and potential mechanisms for human therapeutic interventions and prophylaxes. A number of possible microbiome-based interventions have been proposed in the current scientific literature. Probiotics manipulate the microbiome community through the administration of live microorganisms indented to supplant negative bacterial species or to enhance the abundance of beneficial species in the microbiome (Kobyliak et al. 2016, Yoo and Kim 2016). Prebiotics influence the microbiome through supplementation of the host's diet with materials indigestible by the host but fermentable by the bacteria in the gut (Varankovich et al. 2015, Cockburn and Koropatkin 2016). In fecal microbiome transplants, the microbiome from a presumed healthy individual is transferred into a new host (Khoruts and Sadowsky 2016, Marotz and Zarrinpar 2016). Already, microbiome transplants have proven a powerful tool for curing otherwise intractable diseases such as IBS (Aroniadis and Brandt 2014, Distrutti et al. 2016, Gupta et al. 2016) or antibiotic resistant *Clostridium differences* infections (Brandt et al. 2011, Borody et al. 2013, Rubin et al. 2013, Gupta et al. 2016). Perhaps the most direct method of manipulating the microbiome community is through host diet (Muegge et al. 2011, Wu et al. 2011, David et al. 2014b). Diet provides fuel for both the host and the microbiome, but diet can also act to influence the microbiome in more subtle ways. Sources of nitrogen, polyphenols, triglycerides, and flavanoids in the diet have been shown to have an effect on microbiome community dynamics (Amiot et al. 2016, Huang et al. 2016, Liu et al. 2016, Portune et al. 2016, Rial et al. 2016, Thaiss et al. 2016, Sung et al. 2017).

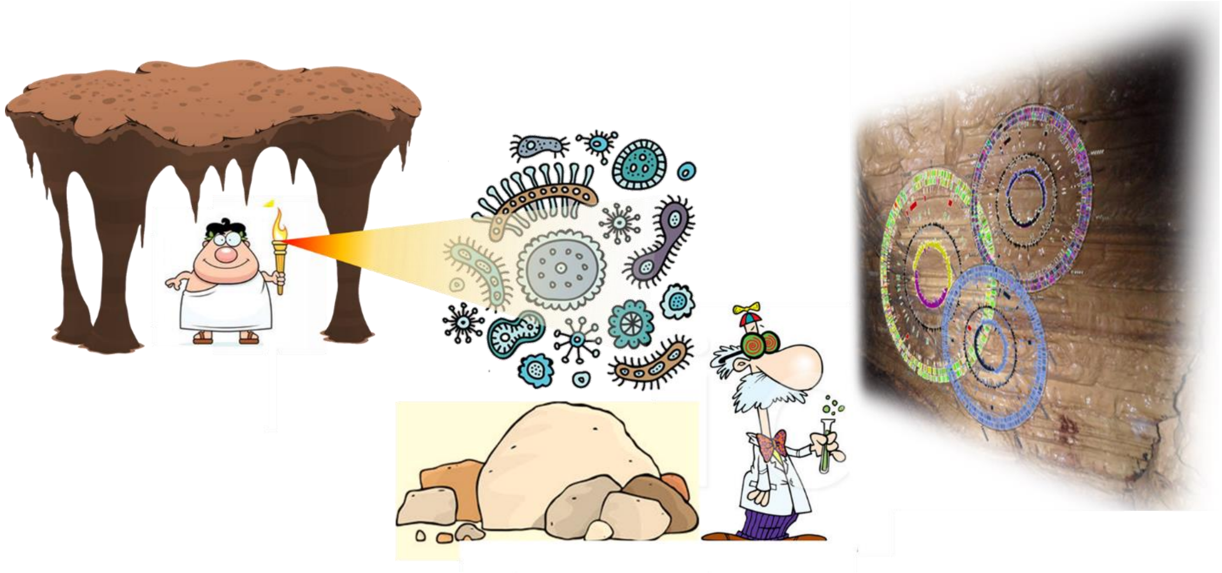


Figure 2.1. Plato's microbiome. It is important to remember that our view of the microbiome is always through some intermediate form of observation. We do see not the microbiome itself, but the molecular shadows cast by the microbiome.

2.3 Methods for Investigating the Microbiome

While the microbiome undeniably plays a crucial role in human health, it is important to remember that no one has ever actually seen a microbiome *in corpus*, as it were, in the flesh. The vast majority of microorganisms in the microbiome cannot be easily cultured in the laboratory as a single strain and, for the most part, one bacteria looks pretty much like another under the microscope. To study the microbiome, we must resort to indirect methods of observation, sorting through extracted nucleotides or proteins or metabolites culled in bulk from microbiomes, then re-assembled in order to reconstruct a picture of the intact community (**Figure 2.2**).

2.3.1 Metagenomics.

Perhaps the most common method for studying the microbiome is through metagenomics DNA sequencing approaches (Gilbert and Dupont 2011, Gjesing and Pedersen 2012, Larsen et al. 2012b) . In metagenomics, DNA is extracted directly from complex communities, sequenced, and analyzed. The resulting DNA sequence data is use to generate a reconstructed image of the microbiome. There are two principle methods for analyzing microbiomes from sequenced DNA: 16S rRNA sequence analysis and shotgun metagenomics.

2.3.1.1 16S rRNA Sequence analysis.

The ribosome is a crucial component of the molecular machinery for translating RNA sequences into protein. In bacteria, a key component of the ribosome is the 30S small subunit of a prokaryotic ribosome (16S rRNA). Due to its essentiality for living processes, all bacteria have at least one 16S rRNA gene. Due to the specificity of function of 16S rRNA, it has a slow and predictable rate for the accumulation of mutations. These characteristics have made it possible to utilize the 16S rRNA ribosomal subunit as a measure of bacterial phylogeny (Gilbert and Dupont 2011, Larsen et al. 2012d). In this analysis, the 16S rRNA genes are sequenced from PCR fragments designed to recognize invariant regions of the gene. As the vast majority of bacteria in the gut microbiome (and, in fact the world of bacterial ecology in general) the most commonly accepted measure of taxonomy using 16S rRNA data is the Operational Taxonomic Unit (OTU). An OTU is group of 16S rRNA sequences within a microbiome community, typically at a sequence identity of 97%, that is presumed to represent a single bacterial species (Gilbert and Dupont 2011). Often, the identity of an observed OTU is determined by closed homology to 16S rRNA gene in previously sequenced and annotated bacterial genomes.

A number of publically available tools, such as QIIME (Caporaso et al. 2010) and muther (Schloss et al. 2009) analyze 16S RNA sequence data, generating microbiome community structure data from UniFrac distances (Lozupone et al. 2007). Other tools generate microbiome community structure from metagenomics sequence data. The principle advantage of 16S rRNA data is the relative ease at which this form of data can be collected and analyzed. The main disadvantages in using 16S rRNA sequence data is that the main assumptions are sometimes violated in observed sequenced genomes: a bacterium may have more than one copy of 16Sr RNA gene per genomes and bacteria with similar 16S rRNA genes sequences sometimes are quite divergent in the rest of their genomic sequences. Also, the selection of specific PCR primers used in analysis can contribute to bias in the final predicted microbiome community n structure.

2.3.1.2 Shotgun metagenomics

In shotgun metagenomics, all of the DNA present in a microbiome is sequenced (Gilbert and Dupont 2011, Gilbert et al. 2011, Larsen et al. 2012b, Noecker et al. 2017). The results of sequencing are DNA fragments that can be computationally assembled into longer stretches of contiguous sequences (contigs) up to potentially complete bacterial chromosomes. Contigs are searched for probable gene sequences and the functions of potential genes are determined by homology to genes/proteins that have previously been assigned functional annotations. The advantages of this method is a far more complete picture of bacteria and distribution of genes present in a microbiome that can be collected from 16S rRNA data alone. The disadvantages include that shotgun metagenomics is far more expensive than 16S rRNA sequence analysis, both in terms of sequencing costs and computational effort. Assembly of contigs can also lead to chimaeras, erroneous computational assembly of fragment in into contigs that do not represent DNA sequences actually present in the population. Annotation of shotgun metagenomics can also be subject to error as, in some cases, even high sequence homology between genes does not always translate into genes of similar function and the existing databases of protein functions can be subject to incorrect or irrelevant annotations.

Prediction of microbiome community structure from metagenomic sequence data is a potentially more accurate approach than 16S rRNA analysis as metagenomic sequences are presumably free from the sequence bias potentially introduced by 16S rRNA amplification steps. MEGAN is one tool for identifying the taxonomic content of a microbiome community from its sequenced metagenome (Huson et al. 2007). MEGAN works by aligning metagenomics sequences to existing databases, such as by BLAST, then utilizing the database of NCBI taxonomy to rank the results. Reads are assigned to a taxon using a lowest common ancestor algorithm. CARMA is an alternate approach that uses a search for conserved functional protein domains with a set of translated DNA sequences and assembles a phylogenetic tree based on protein function similarities, and classified reads into higher-order taxonomy (Krause et al. 2008). MetaPhyler is an approach that identifies a microbiome's community structure from metagenomic data by searching for the presence of specific marker genes in the metagenome that can be uniquely ascribed to a specific taxonomy with high confidence, although the overall accuracy of this approach is limited by the ability to correctly identify suitable marker genes (Liu et al. 2011). Kraken addresses a particular challenge in determining microbiome community structures from metagenomic sequence data which is the speed of application (Wood and Salzberg 2014). Kraken uses k-mer exact alignments to substantially speed microbiome community analysis and achieves accuracy comparable to that of the BLAST algorithm.

In addition to using metagenomic data to determine microbiome community structure, metagenomics data can also be used to identify the specific nature, function, and relative abundance of genes present in a microbiome community. Gene function and identify is determined by sequence homology to databases of genes and proteins of known function. MG-RAST (Meyer et al. 2008, Keegan et al. 2016), and Integrated Microbial Genomes (IMG) (Chen et al. 2017) are commonly available tools for annotating metagenome sequence data with function. The accuracy of metagenome annotation approach is necessarily limited to the quality of the reference database

It is also possible to infer the metagenomic sequence annotation directly from microbiome community structure. PI-CRUST generates prediction of a microbiome's metagenome from community structure by assigning a detected OTU to the most similar bacteria with a completely sequenced genome according to 16S rRNA sequence homology (Langille et al. 2013). Tax4Fun elaborates on this approach by using a linear combination of precomputed genomic reference profiles instead of simply the available genome identified by 16S rRNA gene homology (Asshauer et al. 2015). Potential weaknesses of the PI-CRUST and Tax4Fun methods is the assumption that the closest 16S rRNA homology is always the best sequenced genome for representing an OTU and the lack of the ability to optimize metagenome predictions given a set of known metagenomes.

A new approach that specifically addresses these weaknesses, Taxonomic Average function Profile prediction (TAP-prediction), is described in Chapter 3 and elaborated in Chapter 6.

2.3.2 Metatranscriptomics

Metatranscriptomics is the collection and sequencing of the message RNA (mRNA) present in a microbiome community (Gilbert and Dupont 2011, Gilbert and Hughes 2011, Gilbert et al. 2011, Franzosa et al. 2014, Mandal et al. 2015, Bashiardes et al. 2016). Similar to shotgun metagenomics, metatranscriptomics looks into the molecular functions present in a community of organisms that is not or cannot be cultured, but rather than DNA, the functional potential of a community, metatranscriptomics considers the mRNA, the functional actualization of a community. While the data analysis approaches are generally similar to that for shotgun metagenomics (i.e. sequence assembly, gene predictions, and annotation of the assembled gene sequences), the metatranscriptome is in many ways more informative than the metagenome. The transcriptome identifies populations that are transcriptionally active and those bacteria that are actively undergoing growth and metabolism, as opposed to those that may be present by their DNA sequence but are not contributing to the metabolomic capacity of the microbiome community.

*2.3.3 Metaproteomics***.**

Metaproteomics is the collection and identification of the proteins present in a microbiome. Theoretically, proteomics is the analysis approach that comes closest to identifying the true functional capacity of a microbiome community (Verberkmoes et al. 2009, Xiong et al. 2015). Proteomics identifies not only what proteins are potentially present (as in metagenomics) and are currently expressed at some level in the community (as in metatranscriptomics), but also potentially post-translational modification to proteins and the rates of degradation of proteins (Haange and Jehmlich 2016). There are however currently significant limitation to the capacity to perform proteomic analyses of microbiomes. The cost of proteomics is high in terms of time, money, and computation and the coverage of metaproteomic analysis is low, with only a small fraction of the possible proteins potentially present in the gut microbiome that can be confidently identified.

*2.3.4 Metabolomic***s**

Community Metabolomics (perhaps more appropriately, but less commonly used ‘meta-metabolomics’ (van Baarlen et al. 2013)) is the analysis of the small molecules in a community that are the end products of bacterial processes. Modern metabolomics approaches can detect thousands of small molecules in biological samples through nuclear resonance spectroscopy and mass spectrometry. (Meta)Metabolomics has the capacity to see past the vast diversity of microbiome community composition and multiplicity of enzyme functions to identify the microbiome’s ‘metabolomic phenotype’ that is a key emergent property of the microbiome’s community driving HMIs (Ursell et al. 2014, Smirnov et al. 2016, Shaffer et al. 2017).

*2.3.5 Multi-omic*s

Ideally, microbiome analyses draw from multiple ‘omics data types (Turnbaugh and Gordon 2008, Jansson et al. 2011, Larsen et al. 2012d, van Baarlen et al. 2013, Larsen et al. 2015b). By combining approaches, these methods for investigating the microbiome supplement one another and identify interactions across scales and biological data types that would be invisible by any other means.

2.3.6 Computational modeling

The final component to the analysis of microbiomes is computational modeling (Henry et al. 2011, Larsen et al. 2012b, Larsen et al. 2012d). A computational model of the host-microbiome interactions should, ideally, have the capacity to utilize multiple types of microbiome omics data, detect all possible mechanisms of interactions within the microbiome community and between the microbiome and the host. These models should be able to propose the specific molecular mechanisms of HMI and lead to relevant biological experiments to validate model predictions. The ultimate goal of modeling is to generate potential therapeutic interventions to rationally manipulate host-microbiome interactions. Computational modeling is a kind of lens that permits us to peer into those biological regions not otherwise amenable to direct observation and make meaningful discoveries about what is happening there. The ultimate goal of computational modeling is to provide a bridge from data analysis and machine learning back to experimental biology where it can drive deeper understanding of HMI and propose specific molecular mechanisms of HMI.

2.3.6.1 Modeling Microbiome Community Metabolism

A common approach for generating metabolic models from annotated genomes or metagenomes is Flux Balance Analysis (FBA). FBA is used to computationally simulate growth of an organism is response to a chemically defined environment (Varma and Palsson 1994a, b). In FBA, the interior of the modeled cell is assumed to exist in a quasi-steady state and transport reactions across the cell membrane are only allows if the transported ligand is in the set of defined environmental parameters. ModelSeed (Henry et al. 2010) and KBase (https://www.kbase.us/) are examples of automated platforms for generation genomic or metagenomics FBA models.

One potential weakness of FBA models is its dependence on a defined media environment and the typically used assumption that FBA models should be solved to optimize biomass. Also, FBA are less able to predict the behavior of secondary metabolites that are not directly associated with biomass accumulation. An alternative approach, Predicted Relative Metabolic Turnover (Larsen et al. 2011), is described in in Chapter 3.

2.3.6.2 Models for Predicting Microbiome Community Dynamics

There are also a variety of computational tools that predict microbiome community composition or population dynamics from environmental parameters (Larsen et al. 2012b, Larsen et al. 2012d). These methods differ in the kinds of data used to make a prediction and in the specific nature of microbiome predictions attempted.

Bioclimatic models: Bioclimatic models link microbiome community and presence and absence of possible bacterial species to environmental parameters. One bioclimatic modeling approach is to generate networks of correlative interactions between abiotic environmental conditions and biotic measurements. Bioclimatic models delineate the potentially habitable ranges of a species as a function of environmental parameters (Risto K. Heikkinen 2006, Heikkinen et al. 2007, Jeschke and Strayer 2008). Envelope models, species distribution, and ecological niche model are all examples of bioclimatic models. Multiple computational approaches can be used to generate bioclimatic models, including generalized additive models (Hastie and Tibshirani 1990, Hastie et al. 1992), logistic regression (Bolker et al. 2009), classification and regression trees (Che et al. 2011), fitting of the minimal envelope that defines an bacterium potential habitats in high-dimensional parameter space (i.e., BIOCLIM (Busby 1991), DOMAIN (Carpenter et al. 1993), and HABITAT (Walker and Cocks 1991)); and neural networks (Stockwell and Noble 1992, Stockwell and Peters 1999, Larsen et al. 2012a, Larsen et al. 2012c).

Function-based models: Functional models parameterize bacterial metabolic capacity to generate ecosystem models, linking the environmental metabolome with the environmental conditions. Functional models use factors that represent the aggregate activities or functional capacities for groups of multiple bacterial species (Hood et al. 2006, Ward et al. 2010). In diversity-based models the interactions between environmental conditions and specific bacterial functional traits are modeled (Bruggeman and Kooijman 2007, Follows et al. 2007, Merico et al. 2009). Bacterial functions include biological features like cell size, growth rate, or capacity to metabolic specific nutrients potentially found in the environment.

Individual-based models: Individual-based (IB) models link the bacterial community structure and the community metabolome to predict microbiome dynamics. IB represents a microbiome by creating a representation in the model for every individual cell and that cell's metabolism in the community and representing the environment as a lattice where the nutrient composition for each cell in the lattice is defined (O'Donnell et al. 2007, Ferrer et al. 2008). Since IB models are computationally intense, this approach is best suited to small areas, brief time scales, and relatively small numbers of well-characterized bacteria (Scheffer et al. 1995, Gras et al. 2010).

2.3.6.3 Computationally Modeling the Relationships between the Microbiome and Human Disease

One previously published and highly relevant attempt to link community metabolome with HMI has been described by Shoaie *et al*. They describe a computational approach, i.e. Community and Systems-level Interactive Optimization (CASINO) (Shoaie et al. 2015), for linking modeled microbiome community dynamics to host phenotype. CASINO models the human microbiome as a mix of a selection of culture-able bacterial species representative bacteria commonly in high abundance in the gut microbiome (e.g. *B. thetaiotomicron, E. ractale, B. adolescentis, F. praumsitzii,* and *L. reuteri*). Highly detailed FBA models of communities, optimized to biomass and constrained to the relative abundance of each species were constructed. Models were constructed in two variations: as metabolisms compartmentalized by bacterial species and a mixed-bag approach considering microbiome metabolome as a single metabolic model comprised of the combined metabolic activities on microbiome community model species using the following assumptions: host diet was converted into bacterial growth conditions considering three short chain fatty acids (SCFAs) and fourteen amino acids, and assuming that all host's dietary carbohydrates are converted entirely to glucose molecules. The host's phenotype in this system was defined as the blood serum and fecal composition for SCFAs and amino acids, a subset of which correlated with several observed host obesity-related phenotypes. The model was constructed on parameters derived from laboratory cultured in M2 glucose media and then validated using diets and microbiome community compositions from 'overweight' and 'obese' human subjects and the model predicted some metabolite levels in feces. The CASINO model was used to determine what changes in diet in Low Gene Counts (LGC) individuals, specifically an increase in eight amino acids in the diet, would modify the microbiome community metabolism to more closely resemble High Gene Count (HGC) microbiome metabolisms.

While the CASINO model makes a number of interesting observations and, in some cases, generates a model that can correctly reproduces some aspects of biological observations, it has some notable limitations. The CASINO model could determine changes in microbiome metabolism as a function of the host's diet parameters, however, it does not determine changes in microbiome community structure. Serum amino acid levels, while important diagnostic indicators in a variety of possible disease sates including obesity-related phenomenon, are not necessarily directly a direct function of the digestive processes in the gut microbiome and the diagnostic markers also derive from other biological phenomenon including endocrine disorders, liver diseases, muscle diseases, neoplastic diseases, neurological disorders, nutritional disturbances, or renal failure (Hortin 2012). It is also unclear if a small number of bacteria can effectively model the full complexity of diverse microbiome communities and that the behavior of bacteria cultured alone in minimal media closely mimics the dynamic behavior of bacteria in complex communities and in the highly variable environment of the human gastrointestinal tract.

3. CURRENT STRATEGIES FOR MICROBIOME ANALYSIS AND COMPUTATIONAL MODELING

In our previous and current research efforts, we have developed a diverse collection of microbiome analysis tools. The following descriptions outline our analysis tool set that have been utilized in this study, but have not necessarily been developed directly in the course of this analysis.

First, we will define the terms used throughout this analysis. The purpose of this is to avoid any ambiguity in terms used to describe microbiome characteristics, features, and datatypes. Then we will briefly outline the computational tools that will be used throughout this study, the nature of their required inputs and resultant outputs, and support for these approaches in the published literature.

3. 1 Definitions of Microbiome Terms

An effort has been made to use a consistent terminology for the types of data and microbiome interactions used in this analysis.

3.1.1 Microbiome Community Structure

Microbiome community structure is a vector of the taxonomic identities and relative abundances of bacterial species present in a single microbiome. Generally this refers to a 16S rRNA amplicon dataset, but can also be from analysis of 16S rRNA sequences in a shotgun metagenome dataset. In this study, the community structure is most often presented at a taxonomic level less specific than species or OTU. A microbiome community structure can contain multiple taxonomic levels. A collection of microbiome community structures can be organized as a microbiome community structure matrix, where columns are individual microbiomes and rows are taxon and cells are relative abundances of a taxon in a microbiome population.

While we acknowledge that the full microbiome also potentially includes viral, archaeal, and eukaryotic components of communities, for the purpose of this study, we will use the more common descriptions of microbiomes that are limited to the their bacterial components.

3.1.2 Enzyme Function Profile (EFP)

An Enzyme Function Profile (EFP) is a vector of enzyme function and the relative proportion of the metagenome that is comprised of the genes annotated with that function (Larsen et al. 2015a, Larsen and Dai 2015). An EFP is the subset of genes present in a metagenome that are annotated to code for proteins with enzymatic functions. A collection of EFPs can be organized into an EFP-matrix, where columns are microbiomes and rows are metabolites. When there is the possibility of confusion, a computationally predicted EFP is distinguished from an EFP that is derived from observed shotgun metagenomics data as a 'predicted EFP' and an 'observed EFP' respectively in the subsequent text.

3.1.3 Microbiome Community Metabolome

The microbiome community metabolome is the aggregate metabolic capacity of all the individual members of a microbiome community. While there are a number of possible methods for computing the community metabolome from metagenomic data (many of which were described in Chapter 2), for the purpose of this study, we model the community metabolome using PRMT-scores (Larsen et al. 2011). In the calculation of PRMT-scores, the community metabolome does not consider compartmentalization of individual bacteria, but rather considers a community metabolome as a single, well-mixed reaction vessel comprised of the metabolically active enzymes of all members of the microbiome. A metabolome can be represented as a set of enzyme-mediated interactions.

While the community metabolome might also refer to direct chemical/physical measurements of metabolites present in a microbiome-containing environmental sample, for this study the 'metabolome' refers only to computational models of metabolism and not direct physical measurements.

3.1.4 Host-Microbiome Interactions (HMIs)

Host-Microbiome Interactions (HMIs) are those interactions within a microbiome community and between the microbiome community and its host. These interactions can be associated with some feature, molecular mechanism, or emergent property of the microbiome. Some examples of HMI mechanisms that are specifically relevant for mammalian gut-microbiome interactions are:

- *Biosynthesis of vitamins and secondary bile acid biosynthesis*. The gut microbiome is a vital source of important vitamins, especially vitamins B and K (Conly and Stein 1992, Degnan et al. 2014). Essential amino acids, particularly lysine and threonine, are synthesized by the gut microbiome (Metges 2000, Metges and Petzke 2005). Secondary bile acids are derived in the mammalian gut from primary bile acids by the enzymatic activity of the microbiome. Secondary bile acids can be a risk factor in incidence of colon cancer, particularly for individuals with a high-fat diet (Ajouz et al. 2014).
- *Biocontrol*. Some host benefits provided by the microbiome include defense against pathogens. There are several mechanisms by which biocontrol can occur. Mechanisms of biocontrol may be direct. For example, pathogens may be controlled by the synthesis of compounds with local antibiotic activities or chemo-repellent properties or through predation. Mechanisms may also be indirect, for example, by outcompeting potential pathogens for nutrients or by displacing competing bacterial species for available space or through biofilm formation. The microbiome

can also indirectly contribute to pathogen resistance by interacting with the host immune system and reducing inflammatory responses (Round and Mazmanian 2009, Rooks and Garrett 2016, Shi et al. 2017)

- *Host signaling*. In addition to interactions with host immune systems, the microbiome can synthesize neural and endocrine compounds that interact with host regulatory systems (Foster and McVey Neufeld 2013, Yang and Jobin 2014, Carabotti et al. 2015).
- *Increased access to nutrients in diet*. The gut microbiome allows the host to utilize nutrient sources, such as recalcitrant carbon sources like some plant polysaccharides. The microbiome can also allow the host to extract greater energy from diet, by breaking down fatty acids or complex carbohydrates into simple sugars more readily absorbed and metabolized by the host (Turnbaugh et al. 2006, Rosenbaum et al. 2015).

3.1.5 Emergent Properties

Emergence has been given a number of possible definitions since it was first introduced in the context of the evolution of the human mind in the 19th Century (Corning 2012). For the analyses performed here, we define emergent properties as the interaction of components at different biological scales to create the observable properties in the integrated system that are not present in any of the component elements. We will consider integrated computational models to possess emergent properties when the model can successfully predict known biological behaviors that were not used to train or build the model's component elements.

3.2 A review of previously published analysis tools

A number of previously published computational approaches developed by this laboratory are key elements of the current modeling work. Although not explicitly developed in the course of the work in this thesis, these tools play a prominent role here. These approaches are described briefly in the following sections.

3.2.1 Generating community interaction networks (CINs)

A Community Interaction Network (CIN) is a network indicating predicted causal relationship between changes in environmental parameters and changes in bacterial abundances in a microbiome population (Larsen et al. 2012c, Lax et al. 2014, Metcalf et al. 2016). Interactions can be between environmental parameters and a bacterium, or between two bacteria. In these networks, bacteria are frequently represented at some taxonomic level higher than that of species or OTU.

As it is not possible to anticipate all of the possible mechanisms of community interaction before-hand, even if armed with a complete set of sequenced and annotated genomes for all community members, we utilize a statistical approach to describe the web of interactions among a microbiome community, and between a microbiome community and its environment. Communities of bacteria and microorganisms interact with their environment and with one another to form complex networks of positive or negative, mutualistic or antagonistic relationships. Some of those possible interactions are:

- *Syntrophy*: A community must have the basic ability to consume a selected carbon source, either within a single bacterium or divided across multiple species (Morris et al. 2013). Syntrophy may be species dependent (the same set of species always co-occur) or functionally dependent (multiple species may mix and match provided that key metabolic functions are present in the community), or there is a single core species that may form exclusive or non-exclusive partnerships with multiple possible companion species.
- *Colonizers*: Some species may be ‘pioneer’ species and are potentially independent of a specific environmental condition but are required for the subsequent colonization of carbon-specific consortia (Jefferson 2004, Zhang et al. 2006). Colonization may be due to the generation of

extracellular matrixes that permit other species to colonize, breaking down recalcitrant forms of nutrients into forms more easily consumed by other species, or production of chemo-attractants that recruit other species to an environment.

- *Defenders*: The role of some species in microbiome communities is to prevent other species from colonizing (Clay 2014). Possible mechanisms include sequestering carbon making it unavailable to other bacteria, or generation of antibiotics or other forms of chemo-repellants.
- *Microenvironment*: Some species principle role may be to generate a habitable microenvironment for other species in a fashion other than Syntrophy (Wong et al. 2016). For example, a species may be able to generate an anoxic microenvironment in an otherwise oxygen-available condition, provide an electron source/sink that is distinct from enzymes specific for carbon source degradation, or metal scavenging from an environment to make a microenvironment favorable for some bacteria or unfavorable to others.

A number of microbiome community interaction models (Stein et al. 2013, Cardona et al. 2016, Henry et al. 2016) consider only metabolic interactions within and among community members. These approaches, at best, can only capture syntrophic interactions and competition for biomass accumulation.

We utilize a Dynamic Bayesian Network (DBN) approach to generate a Community Interaction Network (CIN) (**Figure 3.1**). Analysis of a CIN topology can provide insight into the relationships between microbiome sub-populations with specific environmental parameters, identify bacterial taxa that interact with changing environmental conditions mostly or exclusively through interactions with other bacterial taxa, the direction and rate at which microbial taxa may migrate from one sub-location to another in a dynamic environment, and propose trophic interactions. While a statistical analysis approach will not necessarily immediately identify the specific nature of an interaction, any and all types of the interactions described above can be captured by CIN analysis. We have been successful in applying this approach in the analysis of several microbiome communities (Larsen et al. 2012c, Lax et al. 2014, Metcalf et al. 2016, Lax et al. 2017).

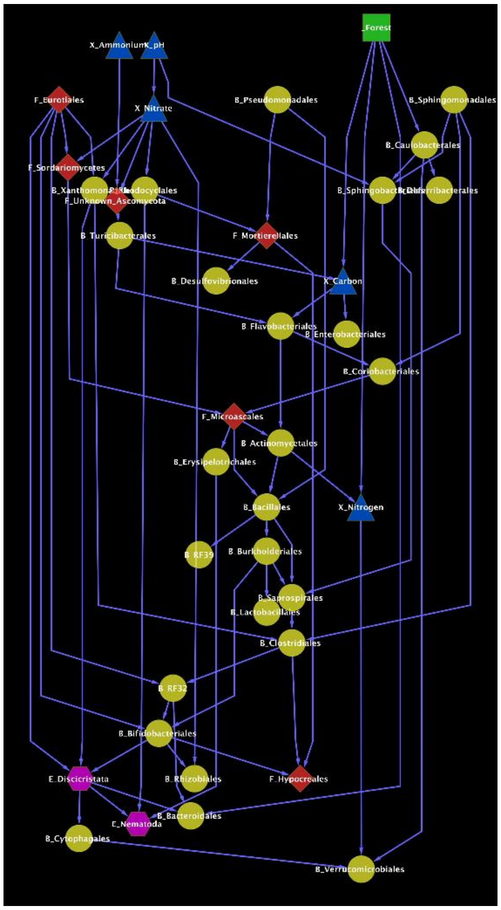


Figure 3.1. Community Interaction Networks (CIN). One key for describing microbiome communities and their interactions with their environment and their host is the CIN. In a CIN, nodes are environmental parameters or microbiome community members. Edges are predicted causal interactions between nodes, as from DBN. The example here is from a previous publication (Metcalf, 2015) for the analysis of how microbiome community members change in abundance in relationship to changing soil parameters during mammalian corpse decomposition.

3.2.2 Microbial Assemblage Prediction (MAP) models

While a CIN may implicitly contain all of the possible interaction types described above, the CIN itself it does not identify their specific nature of interaction, or suggest the directionality (i.e. positive or negative) of the interaction. A number of microbiome community models consider only syntrophic interactions, modeling communities as sets of individual metabolic pathways linked to their environment and to one another through suites of transmembrane transporters (Cardona et al. 2016, Henry et al. 2016). While powerful and useful in their own right, metabolism and a presumed goal of optimizing biomass alone cannot account for most of the key positive and negative interaction mechanisms that drive microbiomes. Synthesis of chemo-attractants that recruit organisms into a community may indeed increase the biomass of bacterial taxa within a community, but the mechanisms of action is communication and regulation, not incorporation of the compound as fuel for biomass in a metabolic network.

In previous publications, we have proposed a method for modeling microbiome communities using an approach that is called Microbial Assemblage Prediction (MAP) models (Larsen et al. 2012a, Larsen et al. 2012c, Larsen et al. 2015b). The first step in this modeling approach is to generate a CIN as described above in 'Community Interaction Networks'. To transform the network structure of the CIN into a computational engine suitable for generating biological predictions, we use the CIN as a scaffold for a system of equations (**Figure 3.2**). In this system of equations, we can define the value of any node in the network, equivalent to the relative or absolute abundance of a bacterial taxa, as a function of the values of its parent nodes, equivalent to the abundance of other bacterial taxa in the community, the abundance of bacterial taxa at a previous time point, or the measured or inferred values of environmental parameters. Environmental parameters can be nutrient availability (e.g. presence of sugars or other carbon sources), availability and nature of nitrogen-, phosphorus-, or sulfur-containing carbon compounds, metal ion availability, or non-nutritive environmental factors such as photosynthetically available radiation, temperature, and pH, or presence of molecules that have little contribution to metabolism, but crucial effects on cell-signaling, such as bacterial quorum-sensing compounds, alarm pheromones, biofilm formation/dissolution signals, or information-rich compounds derived from a microbiome's host organism/environment. In the context of this study, environmental factors are most frequently defined as the diet parameters of the microbiome's host.

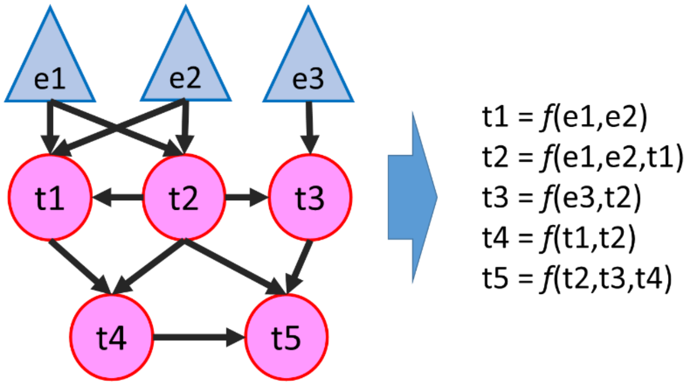


Figure 3.2. MAP-model for predicting microbiome community dynamics. In simplified CIN here, a microbiome community can be described as interactions between three environmental parameters (e1-3) and five bacterial taxa (t1-t5). Network on left can be converted into MAP-model on right be describing network as set of equations for which the value of every node in the network is a function of the values of its parent nodes.

3.2.3 Taxonomic Average Profile (TAP) prediction

The gold standard for microbiome analysis is shotgun metagenomic sequencing, for which the total microbial DNA is extracted from an environment and sequenced, partially or completely assembled into contigs or even full bacterial genomes, and annotated for potential gene regions and their likely functions. While shotgun genomics provides some of the most complete information about a metagenomic community, it is costly and computationally burdensome (Thomas et al. 2012, Jovel et al. 2016). Alternatively, community structure data, in the form of 16S rRNA data is comparatively simple to collect and far cheaper to sequence. This approach focusses only on a single ubiquitous microbial gene, the 16S rRNA, and uses this sequence to propose the evolutionary distance between members of the microbiome community and associate a specific 16S rRNA sequence with the closest available bacterial genome with a sequenced genome or the presence of closely similar 16S rRNA sequences discovered in other metagenome sequence data (described in Chapter 2).

In order to maximize the utility of 16S rRNA microbiome community data, one possible strategy is to leverage previously sequenced and annotated microbial genomes to infer a full metagenome from community structure data. If all of the bacteria present in a community have fully sequenced genomes, then the ability to determine the metagenome as a linear function of sequenced genomes would be a trivial matter, but in the vast majority of metagenomic communities, the community members do not have sequenced genomes, and might in fact be completely uncharacterized by any biological means and known only by its DNA sequences from metagenomic analyses.

One possible method to approach the problem of unknown genomes present in a microbiome community is to find the nearest evolutionary similar bacteria with known genome, inferred by homology with 16S rRNA sequence, and assuming the genome of the unknown bacteria is identical to the sequenced bacteria. This approach is taken by a number of existing tools, such as PiCRUST, SILVAngs, and Tax4Fun (Langille et al. 2013, Quast et al. 2013, Asshauer et al. 2015). The advantage of this approach is that it quickly leverages existing data to infer metagenomic sequences and relies on the assumption that closely-related bacteria will have similar genomes. The disadvantage is that the closest evolutionary neighbor by in the database inferred from 16S rRNA sequence may not be biologically or functionally similar. Even genomes from closely-related bacteria within the same taxonomic group might differ significantly in their actual genomic sequence or possess very different functional capacities.

We have developed and previously published a variation of the linear combination of sequenced genomes approach that provides significant improvement over other metagenome prediction tools. The approach also generates predictions that are ideally suited for MAP-model community predictions and for use in PRMT microbiome metabolome predictions (Larsen et al. 2011, Larsen et al. 2012a, Larsen et al. 2015b). In this approach, Taxonomic Averaged Profile Prediction (TAP-prediction), rather than associate an OTU with the most closely related sequenced genome by 16S rRNA sequence, bacteria are considered at a higher level of taxonomic resolution (e.g. Genus, Order, or Class). The approach is summarized as following:

$$EC_i^n = \sum_{j=1}^{\boldsymbol{Taxa}} TAP(i,j) * Taxa_j^n \qquad \textbf{Eq. 3.1}$$

Where EC_i^n is the abundance of enzyme function *i* in microbiome *n*, ***Taxa*** is the set of bacterial taxa reported present in the microbiome, *TAP(i,j)* is the TAP-matrix for enzyme function *i* in taxa *j*, and $Taxa_j^n$ is the relative abundance of taxa *j* in microbiome *n*. The values of $Taxa_j^n$ are derived from an analysis of thousands to tens-of thousands published sequenced and annotated genomes. A single taxa might be represented from anywhere from very few to hundreds of genomes, and a single microbiome might be described from a dozen or so taxa to hundreds of taxa. The complete set of all $Taxa_j^n$ for all taxa *j* and enzyme functions *i* is called the Taxonomic Average Profile Matrix (TAP-matrix).

The TAP-matrix is a matrix of enzyme functions and the average abundance of genes with those functions in a set of sequenced and annotated genomes collected to represent a specific taxonomic grouping (e.g. Class, Order, or Genus) (Larsen and Dai 2015, Larsen et al. 2015b).

There are several advantages to the TAP-prediction approach. First, it becomes possible to express a predicted metagenome as a function of statistical certainty. While the 'true' metagenome is not known, by incorporating the average and variance of gene function counts in the prediction, it becomes possible to express metagenomic predictions as a statistically-informed range of values rather than a single value representing the expected abundance of an enzyme function in a sequenced genome. The ability to select a taxonomic level for description of prediction metagenomes permits a user to select between tradeoffs for confidence of prediction (i.e. lower taxonomic levels for narrow-distributions of enzyme function predictions with lower confidence, or higher taxonomic levels for wider-distributions of enzyme function predictions with higher confidence). The inputs to TAP-prediction are called TAP-matrix is a collection of vectors of enzyme functions and the average abundance of genes with those functions in a set of sequenced and annotated genomes collected to represent a specific taxonomic grouping (e.g. Class, Order, or Genus). The output of TAP-prediction is an Enzyme Function Profile (EFP), a vector of enzyme functions and their abundances present in a microbiome community. A collection of EFPs is an EFP-matrix.

The TAP-prediction approach is utilized in its previously published form in Chapter 5. In Chapter 6, the TAP-prediction is improved by the incorporation of new statistical analysis approaches that modify the TAP-matrix.

3.2.4 Predicted Relative Metabolic Turnover (PRMT)

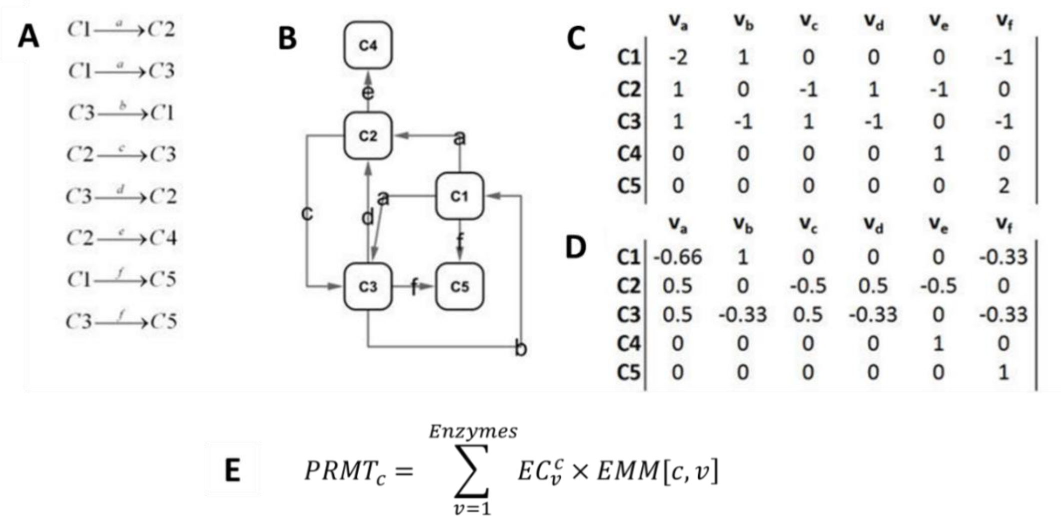
One of the most important features of a microbiome community is the community metabolome, the cumulative metabolic functions of all of its member species modified by their relative abundance. One common approach to inferring a microbiome metabolome from metagenomic data is statistical enrichment/depletion of gene activities found in a metagenome that are annotated to belong to known metabolic pathways, e.g. KEGG pathways. An alternative approach is Flux Balance Analysis (FBA) modeling of microbiome communities (Bucci and Xavier 2014, Henson and Hanly 2014, Bosi et al. 2017). FBA is a computationally inexpensive method for modeling the steady-state metabolic fluxes for genome-scale reconstructions of metabolic networks. Applying FBA to microbiome communities can either consider a mixed model, in which the entire microbiome is modeled as a single 'cell' with a metabolism that is the weighted combination of the metabolic pathways of its constituent members, or else as a set of compartmentalized metabolisms, linked to one another through transmembrane transport reactions.

One disadvantage of this approach is that FBA generally utilizes biomass optimization for generating metabolic models, which relies on metabolic compounds associated with primary metabolism. In many cases, key metabolites associated with microbiome-host interactions are derived from 'secondary' metabolism or with chemical compounds whose primary function in a community is not primarily metabolic but rather environmental signaling and sensing in nature. While FBA is certainly capable of modeling the biosynthesis of such signaling or regulatory compounds, if those compounds are not previously implicated in microbiome community interactions or components of primary metabolism, then those compounds are unlikely to be included in the FBA model. Another disadvantage of this approach is the necessity for incorporating the environmental 'media' composition of the community. While it may be possible to measure the composition of a complex environment such as soil, rumen, or intestine, that information is not always available or easy to collect. Related to this is the Procrustean excess of FBA modeling, which requires that observed data be fit into an existing metabolic network and minimizes the opportunity to let the observations inform the analysis. Finally, in the case of agent-based models, while FBA can be quite rapid, the necessity of calculating potentially thousands of discrete bacterial entities in some environmental space, iterated over any considerable length of time, greatly increases the compute effort of these models, necessitating either simplification of community composition or truncating the 'time' of microbiome community simulations.

To address these limitations when considering microbiome metabolomes, we have developed a novel computational metric, the Predicted Relative Metabolic Turnover (PRMT) score (Larsen et al. 2011). Rather than attempt to create a comprehensive computational model of a single community, PRMT explicitly considers only the predicted relative change in metabolite turnover in one metagenomic dataset relative to another by considering the differences in abundances for enzymes with a specific metabolic activity between metagenomes. While this approach sacrifices the ability to consider the absolute abundances of individual metabolites, PRMT-scores allow all possible metabolic interactions, for primary and secondary metabolism, to be easily calculated. PRMT considers community meta-metabolome as a single, well-mixed reaction vessel comprised of the enzymes with metabolic functions of all members of the microbiome community.

The overall approach for calculating PRMT-score is summarized in **Figure 3.3**. PRMT-scores are unit-less values that represent the change of the turnover of a metabolite in a predicted metabolome relative to a reference metabolome. A PRMT-score is calculated for every metabolite in the EMM for a metagenome. The value and sign (positive or negative) of a PRMT-score provides information about a metabolite's relative turnover. Although a thorough interpretation of a PRMT-score requires that it be considered in the greater context of the complete network, it can be broadly interpreted as follows: A positive PRMT score predicts increased metabolic turnover and relatively greater consumption of a metabolite. A negative PRMT-score predicts decreased turnover and relatively greater accumulation of a metabolite. It is important to note that PRMT-scores do not predict net production or consumption of a metabolite. A metabolite with a positive PRMT is not necessarily being consumed faster than it is being synthesized; only perhaps that it is being synthesized at a lower predicted rate when compared to another metagenome.

The PRMT metabolomic models have been successfully used previously by our lab (Larsen et al. 2011, Larsen et al. 2014, 2015a, Larsen and Dai 2015, Larsen et al. 2015b, Larsen et al. 2016, Shinde et al. 2017) as well as by other researchers (Desai et al. 2012, Mason et al. 2014, Louca et al. 2016).



$C1 \xrightarrow{a} C2$

$C1 \xrightarrow{a} C3$

$C3 \xrightarrow{b} C1$

$C2 \xrightarrow{c} C3$

$C3 \xrightarrow{d} C2$

$C2 \xrightarrow{e} C4$

$C1 \xrightarrow{f} C5$

$C3 \xrightarrow{f} C5$

$$\begin{array}{c|cccccc|} & v_a & v_b & v_c & v_d & v_e & v_f \\ C1 & -2 & 1 & 0 & 0 & 0 & -1 \\ C2 & 1 & 0 & -1 & 1 & -1 & 0 \\ C3 & 1 & -1 & 1 & -1 & 0 & -1 \\ C4 & 0 & 0 & 0 & 0 & 1 & 0 \\ C5 & 0 & 0 & 0 & 0 & 0 & 2 \end{array}$$

$$\begin{array}{c|cccccc|} & v_a & v_b & v_c & v_d & v_e & v_f \\ C1 & -0.66 & 1 & 0 & 0 & 0 & -0.33 \\ C2 & 0.5 & 0 & -0.5 & 0.5 & -0.5 & 0 \\ C3 & 0.5 & -0.33 & 0.5 & -0.33 & 0 & -0.33 \\ C4 & 0 & 0 & 0 & 0 & 1 & 0 \\ C5 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}$$

$$PRMT_c = \sum_{v=1}^{Enzymes} EC_v^c \times EMM[c,v]$$

Figure 3.3. Calculation of community metabolome as PRMT-scores. (A) A set of all enzyme functions from an EFP are collected (a-f in figure). From a set of known metabolic interactions, all possible enzymatic transformations with those enzymes are collected. (B) The enzyme reactions are assembled into a metabolic network. (C) The metabolic network is transformed into a connectivity matrix. (D) Connectivity matrix is normalized such that all in-edges sum to 1 and all out-edges sum to -1. This matrix, for the purpose of this figure, is called the Environmental Metabolic Matrix (EMM). (E) PRMT-scores are calculated for each compound c in the EMM, where EC_m^c is the $\log_2$ normalized relative abundance of enzyme function c in metagenome m and EMM[c,v] is the c^{th} row and v^{th} column of the EMM.

4. PSEUDOMONAS-HOST INTERACTIONS

The ability to assemble complete bacterial genomes from community metagenomics data has become increasingly commonplace (Gilbert and Dupont 2011). However, there may be very little opportunity or capacity to isolate and characterize these bacteria that are known only through metagenomic data. To understand the role that these otherwise uncharacterized bacteria play in their microbiome community, we must turn to computational modeling and machine learning when direct biological characterization of a bacterium is impossible.

The goal of Chapter 4 is to address the ever growing abundance of sequenced bacterial genomes, such as those that might derive from metagenomic sequence from microbiome communities. The metabolomic and transportomic models, constructed from annotated genomes, are used to determine the HMI class filled by a bacterium through application of machine learning techniques. The result of this study is a computational tool for the identification of the specific bacterial molecular mechanisms that drives HMIs. This information can be used to identify ways positive host interactions can be enhanced and negative interactions can be mitigated, such as addressing antibiotic resistance in hospital acquired bacterial infections.

4.1 Background

Pseudomonads, Gram negative bacteria in the *Gammaproteobacteria* Class, are nearly ubiquitous (Silby et al. 2011). These bacteria can be found in ocean waters, in terrestrial soils, living in symbiotic associations with plant roots, and within the GI tracks of mammals (Clarke 1982, Ridgway et al. 1990, Ryan and Heaner 2014). Their adaptation to a wide range of possible ecological niches is made possible by Pseudomonad's wide array of secondary metabolic capacities (Palleroni 1992).

4.1.1 Pseudomonads Can Occupy Multiple HMI Classes

Here, we focus on Pseudomonas for their capacity to occupy two very different but highly important ecological niches: Pseudomonas are bacteria that form symbiotic relationships with terrestrial plant roots (Lugtenberg et al. 2001, Silby et al. 2009) and Pseudomonads are the second most common cause of pneumonia infections in intensive care units (de Bentzmann and Plesiat 2011).

4.1.1.1 Pseudomonads as Plant Growth Promoting (PGP) bacteria

Just like humans, terrestrial plants also possess complex communities of microorganisms that help them access nutrients, defend against disease, and regulate growth (Cook et al. 1995, Frey-Klett et al. 2011, Kurek et al. 2013, Giles et al. 2014, Cumming et al. 2015). Beneficial bacteria in these plant-root associated microbiomes provide PGP effects, enhancing a plant's capacity to acquire biomass, even in sub-optimal growth conditions. In return for their Plant Growth Promoting (PGP) services, the communities of beneficial microorganisms receive carbon, in the form of photosynthetically derived sugars. *Pseudomonas fluorescens and Pseudomonas protegens*, commonly found in the soil and frequently associated with aspen tree roots, are such PGP bacteria (Gottel et al. 2011, Brown et al. 2012b). Aspen are highly relevant trees for study as they are the most abundant tree in the American Northwest and are an important tree species used as an experimental model organism for understanding terrestrial carbon sequestration for the Department of Energy (DOE) (Djerbi et al. 2005, Tuskan et al. 2006). While the beneficial effects of *Pseudomonas fluorescens* on aspen is well established, the molecular mechanisms that underlie PGP effects remain unknown. The goal of this analysis on plant-pseudomonad HMIs is to uncover the specific mechanisms by which *Pseudomonads* enhance aspen growth.

4.1.1.2. Pseudomonas as Antibiotic Resistant Hospital Acquired Infections (HAIs)

From the CDC, "Gram-negative bacteria cause infections including pneumonia, bloodstream infections, wound or surgical site infections, and meningitis in healthcare settings. Gram-negative bacteria are resistant to multiple drugs and are increasingly resistant to most available antibiotics. These bacteria have built-in abilities to find new ways to be resistant and can pass along genetic materials that allow other bacteria to become drug-resistant as well." (CDC 2011). In the 2013 CDC Antibiotic Resistance Threat Report (CDC 2013), *Pseudomonas aeruginosa* that develop multidrug resistance was given a threat level of "Serious" with 13% of all *Pseudomonas aeruginosa* HAIs being multi-drug resistant. To counter the rising threat of deadly *Pseudomonas aeruginosa* infections, the goal of this analysis of human-Pseudomonad HMIs is to identify potentially novel targets for antibiotics or development of methods to mitigate their pathogenicity.

4.1.2 Previous publications for Pseudomonas HMI Class Identification

An approach for elucidating the molecular mechanisms of HMI-class occupation by Pseudomonads in Aspen root from genomic sequence data has been published by Timm at al. (Timm et al. 2015). In that study, the ecological niches under consideration were rhizosphere, found living on or very near the surface of aspen roots, or endosphere, living within the cells of aspen roots. In this approach, 19 genome sequences (4 rhizosphere and 15 endosphere) were collected from bacteria cultured from the rhizosphere and the endosphere of environmentally sampled aspen tree roots. From assembled and annotated genomes, FBA models were constructed for bacteria. FBA identified differenced in predicted primary metabolism between endosphere and rhizosphere Pseudomonads, but in this study there was no opportunity to consider the approach as a predictive tool.

A previously published tool for predicting a bacteria's pathogenic capacity from genomic data, Huang et al., is Path-Based Human Microbe-Disease Association (PBHMDA) (Huang et al. 2017). In that study, known microbe-disease associations were used to train a model based on the Gaussian interaction profile kernel similarity for microbes and diseases. In a 5-fold CV scheme, the approach successfully associated 292 bacteria with 39 diseases, collected from HMDAD database (http://www.cuilab.cn/hmdad).

Another computational approach, published by Suzuki et al. (Suzuki et al. 2014), predicts antibiotic resistance from gene expression profiles. Using laboratory evolution of E. coli to several antibiotics, authors identifies instances of antibiotic cross-resistance as well as instances where increased resistance to one antibiotic lead to increased susceptibility to other antibiotics. Using transcriptomic analysis, authors demonstrate that while the specific mutation leading to antibiotic resistance might vary, the resulting change in patterns of gene expression was consistent for a given resistance phenotype.

4.1.3. Key Knowledge Gaps and Innovations

This Aim fills specific knowledge gaps and puts forth a number of key technical innovations. While we have used of PRMT-method for modeling complex metabolisms successfully in the past, in this aim we develop a companion metric to the PRMT-score, the Predicted Relative Transmembrane Transport (PRTT) score. This metric is not only a novel analytical tool, but modeling the Transportome through PRTT-scores proves to be a highly effective approach for predicting Pseudomonad HMI-classes. Although a number of prior published research efforts described above have generated predictive models of HMI, few of them are capable of proposing the molecular mechanisms that drive HMI or to propose specific experiments to validate predictions. Not only do our tools point directly to specific molecular mechanisms of HMI that can be validated through hypothesis-driven biological experiments, but through collaborators at Argonne National Laboratory, a number of predicted mechanisms for Pseudomonad-plant HMIs have already led to experimentation and prediction validation, as described below and first reported in (Larsen et al. 2014, 2015a).

4.2 Outline of Experimental Approach

In order to construct predictive models that associate a Pseudomonad strain with a specific HMI class, a highly-curated collection of data needed to be acquired. Needed for this analysis were:

- A set of human and aspen Pseudomonad HMI classes that could be confidently assigned to a Pseudomonas species in strict Boolean fashion.
- An available set of sequenced and annotated Pseudomonas genomes.
- A clear, published reference by which a Pseudomonas strain could be ascribed to one of the selected HMI classes.

The set of niche types was taken from a Pseudomonas review published by Silby et al (Silby et al. 2011). The host interaction types selected span plant-related positive and negative PGP effects and human pathogenicity interactions:

- *Antibiotic Resistance*: Bacteria with ‘Antibiotic Resistance’ HMI have the Mobile Genetic Elements (MGEs) and efflux systems that confer resistance not only to antibiotics, but potentially also other stress conditions (Arnold et al. 2003, Hacker et al. 2004, Jackson et al. 2005).
- *Biocontrol*: Pseudomonas can provide PGP functions to its plant host by antagonizing potential pathogens, directly influencing the host’s growth and disease resistance (Haas and Defago 2005, Ryan et al. 2008, Garbeva and de Boer 2009).
- *Biofilm formation*: Biofilms are complex structures comprised of multiple individuals bound together on a substrate through the excretion of materials such as extracellular polysaccharides.

The capacity for forming biofilms includes the capacity to send and receive specific messages to other members of the same bacterial species to coordinate biofilm formation. Biofilms can protect communities against antibiotics and create oxygen limiting microenvironments that influx local redox conditions (Rudrappa et al. 2008, Alhede et al. 2014, Quiles and Humbert 2014).

- *Pathogen*: Pseudomonas have the capacity to cause disease in a wide range of organisms. Pathogenicity includes possession of virulence factors and specific secretion systems. Pathogenicity often co-occurs with biofilm formation and antibiotic resistance (Clarke 1982, Silby et al. 2011, Fernandez et al. 2015).
- *Plant Growth*: Pseudomonas have a variety of mechanisms for inducing increased growth of plants, including the capacity to mobilize nutrients in soil that would otherwise be unavailable to plant roots or stimulation of root growth through production of plant signaling compounds (Silby et al. 2009, Frey-Klett et al. 2011)
- *Plant Disease*: Some Pseudomonads have the ability to cause disease in plants, a very distinct phenotype from the ability to cause disease in animals. Plant disease is influenced by the ability to inject proteins into plant cells and the capacity to synthesize phytotoxins (Sands et al. 1970, Silby et al. 2009, Mansfield et al. 2012).

These HMI classes are non-exclusionary and any Pseudomonad may belong to some, any, all, or none of them. Niche types represent that capacity to occupy a niche, and not the requirement that it do so. Membership to an HMI class by a Pseudomonas can be conditional, for example a single bacterium may be a plant growth promoting under some circumstances and cause plant disease in others, but the capacity for HMI-class membership is inherent to the bacterial strain.

4.3 Selected Pseudomonas Genomes for Analysis

At the time of this research, there were 43 sequenced and annotated Pseudomonas available that could also be traced to a specific HMI class through published references. The Pseudomonad species, number of genomes associated with that species, the membership to ecological niches, and the relevant references are listed in **TABLE 4.1**. Genome references can be found in **Appendix A**.

Table 4.1. Available Pseudomonad Genomes with Known HMI Classes

Species	**# Genomes**	**Human Pathogen**	**Antibiotic Resistance**	**Biocontrol**	**Biofilm**	**Plant Pathogen**	**Plant Growth**	**References**
Aeruginoa	9	Y	Y	N	Y	Y	Y	(Silby, Winstanley et al. 2011)
Brassicacearum	1	N	N	Y	N	N	N	(Ortet, Barakat et al. 2011)
Denitrificans	1	N	N	Y	N	N	N	(Ainala, Somasundar et al. 2013)
Entomophila	1	N	N	Y	N	N	N	(Vodovar, Vallenet et al. 2006)
Flourescens	4	N	Y	Y	Y	Y	Y	(Silby, Winstanley et al. 2011)
Fulva	1	N	N	N	N	N	N	(Renault, Deniel et al. 2007)
Mendocina	2	Y	N	Y	Y	N	N	(Silby, Winstanley et al. 2011)
ND	1	N	N	N	N	Y	N	(Li, Zhao et al. 2013)
Poae	1	N	N	Y	Y	N	Y	(Muller, Zachow et al. 2013)
Protogens	2	N	N	Y	Y	Y	Y	(Jousset, Schuldes et al. 2014)
Putida	11	Y	Y	Y	Y	N	Y	(Silby, Winstanley et al. 2011)
Stutzeri	6	N	Y	N	N	N	Y	(Silby, Winstanley et al. 2011)
Syringae	3	Y	Y	Y	Y	Y	N	(Silby, Winstanley et al. 2011)

4.3.1 Standardize Annotation of Pseudomonad Genomes

To reduce the possibility that differences between Pseudomonas by their annotated genomes are due, in part, to variations in how and when their genomes were annotated, the 43 genomes used here were re-annotated to a uniform set of conditions. To accomplish this, a novel database of annotated protein sequences was generated using the database of Kyoto Encyclopedia of Genes and Genomes (KEGG) (Ogata et al. 1999, Bairoch 2000, Mitra et al. 2011, Kanehisa et al. 2012).

4.3.1.1 Database of Annotated Bacterial Proteins

Two sets of protein annotations were considered in this analysis: Enzyme Commission (EC) annotations (Bairoch 2000) for description of enzyme functions and a subset of KEGG Orthology (KO) annotations (Mao et al. 2005) for description of transmembrane transport functions.

4.3.1.2 Unique Enzyme Function Annotations

EC annotations are numerical classifications of enzyme functions based of the specific chemical reactions they catalyze. EC annotations are comprised of four numbers separated by '.'s, in which each subsequent number is associated with increasingly specific descriptions of enzyme catalyzed reaction (Bairoch 2000). All KEGG bacterial proteins with an enzyme commission (EC) annotation were collected for our database of enzyme functions, and were comprised of 2,605 unique enzyme EC annotations and 754,066 protein sequences.

4.3.1.3 Transmembrane Transporter Function Annotations

KEGG Orthology (KO) annotations are a unique resource associated with the KEGG database. KO annotations link experimental evidence of enzyme function through sequence homology. Similar to EC annotations, KO annotations are grouped into a hierarchical structure of increasing specificity of enzyme function descriptions. For this analysis, the subset of KO annotations from the "Metabolism/Environmental Information Processing/Membrane Transport" category were used. For each annotation, the set of ligands transported by that annotation were manually curated and ligand identifiers were selected to match the ontology of metabolites in KEGG reactions. All KEGG bacterial proteins with a transmembrane transporter KO annotation were collected, resulting in a database of 164,321 protein sequences annotated to one of 891 unique transporter functions and associated with the transport of 272 possible ligands or environmental information type.

4.3.1.4 Re-annotate Pseudomonad Genomes

Amino acid fasta formatted (*.faa) files were collected from the NCBI database (ftp://ftp.ncbi.nlm.nih.gov/genomes/) for all 43 Pseudomonas genomes. Pseudomonad genomes were re-annotated by best BLAST-P (NCBI-Blast 2.2.23+) hit with e-values < $1x10^{-10}$ between predicted Pseudomonas gene product sequences and our generated databases of enzyme and transporters. Note that by this method, it is possible for a single protein sequence to have both an enzyme and transporter annotation. The complete set of re-annotated genome information can be found in the published reference (Larsen et al. 2015a). Genome annotations are in the format of vectors of enzyme or transporter functions and the number of genes in the genome with that annotation. In this way, all annotated genomes can be described by vectors of uniform lengths.

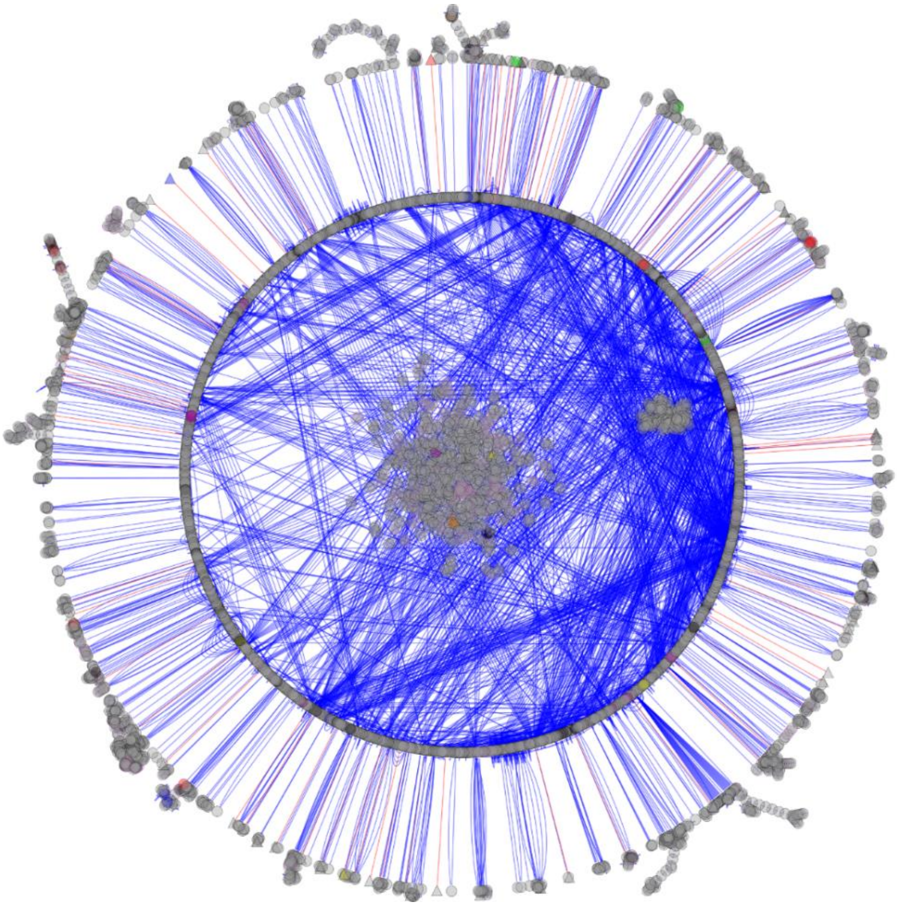


Figure 4.1. Pan-Pseudomonas metabolic and transportomic model. In this visualization, metabolites are circles and extracellular ligands transported by transmembrane transporters are triangle. Enzyme mediated metabolic interactions, transforming one metabolite into another, are blue edges. Transmembrane protein-mediated transport interactions are red edges. Metabolites highlighted with purple edges are secondary metabolites. Nodes are colored by their predicted association with a specific host interaction type: purple for antibiotic resistance, orange for biocontrol, blue for biofilm formation, red for pathogen, yellow for plant disease, green for plant growth promotion, and black for membership to multiple host interaction types.

4.3.2 Generate Metabolomic and Transportomic Models

Using PRMT and PRTT (Larsen et al. 2011, Larsen et al. 2014, 2015a), as described in detail in Chapter 3, metabolic and transportomic models were made, calculated relative to the average enzyme function profile or transporter function profile across all 43 Pseudomonads. A Secondary PRMT metabolome was generated using only enzymatic reactions found in KEGG Pathway 01110, "Biosynthesis of Secondary Metabolites" (http://www.genome.jp/keggbin/show_pathway?map01110).

The resulting pan-Pseudomonas model is visualized in Figure 4.1. The model has 6642 enzyme interactions between 3688 metabolites mediated by 2386 unique enzyme functions and 271 transmembrane transported ligands. Of the metabolic interaction, there are 1649 enzyme mediated interactions between 1494 metabolites associated with Secondary Metabolism.

4.4 Train SVMs to Identify Pseudomonas HMI Classes from Microbiome Data

The re-annotated genomes and metabolic and transportomic model data were used to train Support Vector Machines (SVMs) to predict Pseudomonad HMI classes according to the procedure described below.

4.4.1 Feature Selection for Training SVMs

There are four data types collected or generated for the 42 Pseudomonads: Enzyme Function Profiles, Metabolome and Secondary Metabolome models (as PRMT-scores), and Transportome model (as PRTT-scores).

For all metabolic and transportomic models, training data were restricted by the following conditions:

- Due to the nature of PRMT calculations, some sets metabolites will have identical PRMT-scores across all genomes (e.g. fatty acids of different lengths, in which the same enzyme functions are involved in multiple steps of the fatty acid biosynthesis pathway). All metabolites with identical PRMT-scores were concatenated into a single meta-metabolites.
- All metabolites with PRMT-scores with a standard deviation less than 0.2 across all genomes were removed to subtract metabolites for which there is little variation in metabolic capacity across Pseudomonads.

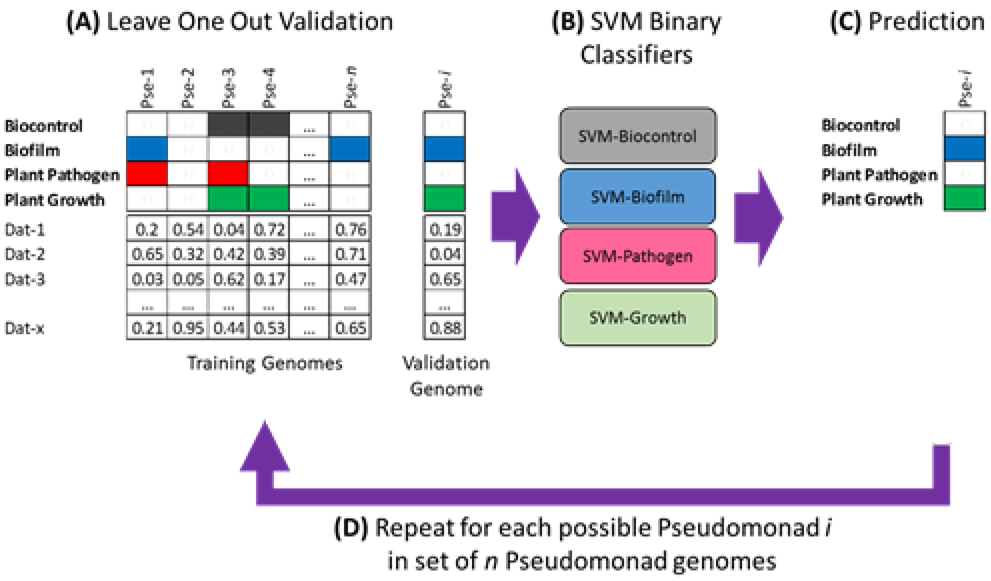
After this down-selection of features, the following number of features, by data type, remained:

- 606 Enzyme Function Profiles (EC annotations)
- 2149 Metabolic Models (PRMT-scores)
- 714 Secondary Metabolism (PRMT-scores)
- 169 Transportome (PRTT-scores)

These selected features were used in the subsequent step for training and validating SVM for prediction of Pseudomonad's host interaction types.

4.4.2 Validate models using a LOO-CV Approach

The four datatypes described above were each used to train Support Vector Machines (SVMs) to predict a Pseudomonad's set of host interaction types. SVMs ia a supervised machine learning approach that, given a novel observatio, predicts its membership to two or more possible classes. Here, SVMs were trainied in a Leave-One-Out Cross Validation (LOO-CV) scheme. LOO-CV is a specific case of an exhaustive leave *n* out cross validation where *n* is equal to 1. One SVM was considered for each of the 6 host interacion types. The outline of the training and validation approach is summarized in **Figure 4.2**. A total of 24 SVM models were constructed (i.e. 6 host interaction types x 4 possible types), which in a LOO-CV scheme totals to 1,032 SVMs (i.e. 24 SVM types x 43 LOO-CV). Construction of SVMs was performed using package 'e1071' in R (R-Project). SVMs were trained using a 10-fold cross validation and linear kernels. In all subsequent discussions, only the results of the validations are presented.


(A) Leave One Out Validation
(B) SVM Binary Classifiers
(C) Prediction
SVM-Biocontrol
SVM-Biofilm
SVM-Pathogen
SVM-Growth
Training Genomes
Validation Genome
(D) Repeat for each possible Pseudomonad *i* in set of *n* Pseudomonad genomes

	Pse-1	Pse-2	Pse-3	Pse-4		Pse-*n*		Pse-*i*
Biocontrol	0	0	⬛	⬛	...	0		0
Biofilm	🟦	0	0	0	...	🟦		🟦
Plant Pathogen	🟥	0	🟥	0	...	0		0
Plant Growth	0	0	🟩	🟩	...	0		🟩
Dat-1	0.2	0.54	0.04	0.72	...	0.76		0.19
Dat-2	0.65	0.32	0.42	0.39	...	0.71		0.04
Dat-3	0.03	0.05	0.62	0.17	...	0.47		0.65
	...	...	...	...	...	...		...
Dat-x	0.21	0.95	0.44	0.53	...	0.65		0.88

	Pse-*i*
Biocontrol	0
Biofilm	🟦
Plant Pathogen	0
Plant Growth	🟩

Figure 4.2. Outline of SVM Training and Validation Method. (A) The Pseudomonas genome data set is comprised of n Pseudomonads. Each Pseudomonad is defined as binary memberships to one of four possible ecological niche types (Biocontrol, Biofilm, Plant Pathogen, or Plant Growth) and an array of x data features. Pseudomonads may be annotated with all, none, or any combination of possible ecological niche types. Pseudomonad genome *i* is selected from the set of n genomes to serve as a validation. (B) SVM binary classifiers are trained on training genomes from (A). Four separate SVM classifiers are generated, one for each environmental niche type. (C) Ecological niche types for validation Pseudomonad *i* are predicted using SVMs from (B). Note that due to use of binary classifiers, it is possible for Pseudomonad i to have a specific pattern of ecological niche type memberships that was not present in the training data set. (D) This procedure is repeated n times such that every Pseudomonad genome is used as validation sample once.

Accuracy of predictions were calculated as F-scores, which is a measure that combines recall and precision. F-score is calculated as

$$F\text{-}score = 2 * (precision + recall) / (precision * recall) \qquad \textbf{Eq. 4.1}$$

$$precision = TP / (TP+FP) \qquad \textbf{Eq. 4.2}$$

$$recall = TP / (TP+FN) \qquad \textbf{Eq. 4.3}$$

Where *TP* is the number of true positives in prediction, *FP* is number of false positives, and *FN* is the number of false negative predictions. An F-score of 1 indicates perfect prediction. The results of LOO-CV are summarized in **Figure 4.3.**

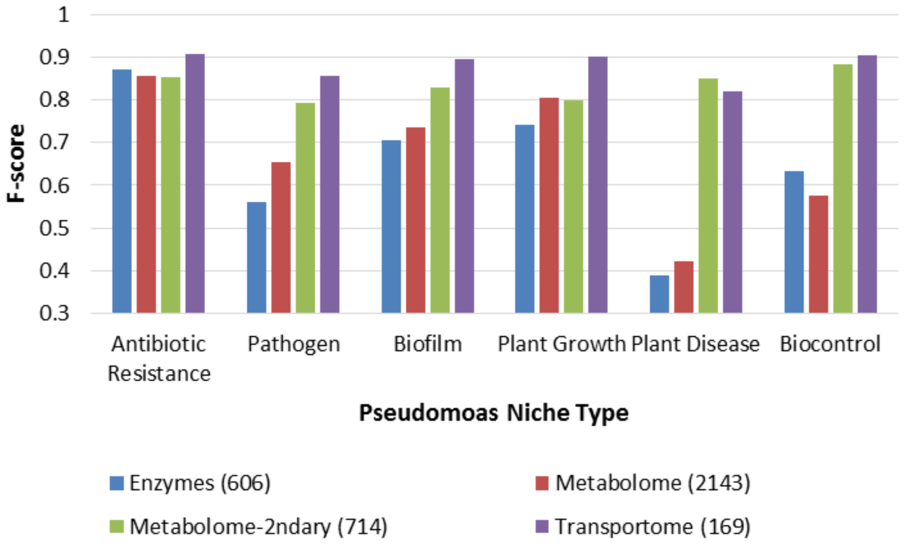


Figure 4.3. F-scores for SVM predictions of host interaction class by data type. The number of features from each data type (Enzyme Function Profile, Metabolome, Secondary Metabolome, and Transportome) are indicated by number in parenthesis. Prediction results are presented as F-score, with an F-score of 1 indicating perfect prediction

For all HMI classes except for Pathogen, the Transportome datatype is most predictive. For all HMI classes except Plant Disease, enzyme function profile is the least predictive. All SVM validation results are relatively strong with F-scores greater than 0.8 for secondary metabolism and Transportome data types. For pathogen, plant disease, and biocontrol, Enzyme Function Profile and Primary Metabolism have the worst F-scores. The predictive power of transportomic and metabolomic data provides an interesting insight into possible mechanisms of HMI. Those HMI that are most strongly associated with interaction with a host and center on sensing and signaling capacity – Pathogenicity of human or plant and Biocontrol – are best predicted by Transportome. Those HMI that might be more likely associated with interaction with specific compounds in the environment –Biofilm formation, Antibiotic resistance, and Plant growth – are best predicted by metabolome.

As Secondary Metabolome based predictions are under all conditions as good or better than predictions based on total Metabolome, only Secondary Metabolism will be considered in subsequent analysis.

4.5 Identify Most Predictive Features for Each HMI Class

While the capacity to predict a Pseudomonad's ability to interact with a plant or animal host is a potentially powerful tool when applied to the analysis of newly sequenced Pseudomonad genomes, prediction of HMI class alone will not lead to novel discoveries into the molecular mechanisms of HMIs. Fortunately, SVM models have a relevant output in addition to class prediction; from SVMs, the relative weight of all features in the predictive model are reported, with higher weight features being more strongly predictive in the SVM. Here, we propose that those features that are most predictive are also most likely to provide biological insight into the molecular mechanisms of host- Pseudomonad interactions.

A feature *i* is identified as highly predictive when:

$$\mathit{FeatureWeight}_i > \mathbf{AVE}(\text{All Feature Weights}) + 2 \text{ x } \mathbf{STDEV} \text{ (All Feature Weights)} \qquad \textbf{Eq. 4.4}$$

Where *FeatureWeight*$_i$ is the reported weight of feature *i*, **AVE** is the average of all feature weights in an SVM and **STDEV** is the standard deviation of all feature weights in an SVM.

Predictive features are listed in **Tables 4.2** and **4.3.**

Table 4.2. Predictive Features of Plant-Microbe Interactions

HMI-Class	Predictive Features	Proposed Molecular Mechanisms
Biocontrol	Cobamide coenzyme (vitamin B12)	Resistance against pathogens and biosynthesis of plant growth factors
	Monosaccharide transport	Plant wound response and pathogen detection
	Acetyl-D-glucosamine metabolism	Defense against fungal infections
	Isoniazid metabolism	Production of antimicrobial compounds
Biofilm	Protoporphyrin and methyglyoxal metabolism	Defense against biofilm formation inhibiting compounds
	Anthranilate degradation	Biofilm formation
	Shikimate pathway	Biofilm formation
Plant Disease	Fatty acid biosynthesis	Lipid signaling in plant-pathogen interactions
	Arabinose and polyamine transport	Plant stress signaling and plant pathogen defense response
Plant Growth Promotion	Indole, eriodictyol, neringenin	Plant signaling compounds association with plant growth promotion
	C4-dicarboxylate transport	Increased organic acid metabolism
	Calcium transport	Plant signaling and associated with plant growth promotion
Rhizosphere	2-O-alpha-manosyl-D-glycerate	Osmoregulation in soil environment
	3-hydroxyphenylpropionic	Capacity to degrade/consume plant material
	Cation transport	Charge balance in soils
	catecholamine biosynthesis	Production of plant stress regulatory compounds

Table 4.3. Predictive Features of Human-Microbe Interactions

HMI-Class	Predictive Features	Proposed Molecular Mechanisms
Pathogenicity	Colicin transport	Bacteriocin
	Homoserine	Signaling compounds associated with biofilm formation
	Sugar/Carbohydrate transport	Virulence factor
	Zinc, MG2+, K+ transport	Highly predictive of pathogenicity in Pseudomonas and associated with metal intoxication
Antibiotic Resistance	Macrolide transport	Antibiotic efflux
	Protein and colicin transport	Bacteriocins are species-specific antimicrobials that are normally produced by Gram-negative bacteria to kill competitors
	Homoserine transport	Signaling/quorum sensing in gram negative bacteria
	Metal ion transport	Metal toxicity resistance is co-selected with antibiotic resistance
	Metabolism of terpenoids and polyketides	Antibiotic degradation pathways
Human Host	3-Hydrosyphenyl-propionate	Affects ROS in tissues and in circulation
	L-Arginine	Immune response to pathogen
	Dopamine	Gut-Brain axis
	Protoporphyrin	Metal complexing in metal poor human environment

The predicted mechanisms for Biofilm Formation, one of the selected Animal-microbe HMI types, is found in **Table 4.2** and is not repeated here.

4.6 Summary of Results

Results have been divided into two biologically meaningful groups: Plant- Pseudomonad Interactions and Human-Pseudomonad Interactions. Plant-Pseudomonad interactions are comprised of Biocontrol, Plant Disease, and Pant Growth Promotion. Human-Pseudomonad interactions are composed of Pathogen and Antibiotic Resistance. Biofilm Formation is a member of both Plant- and Human-Pseudomonad interactions.

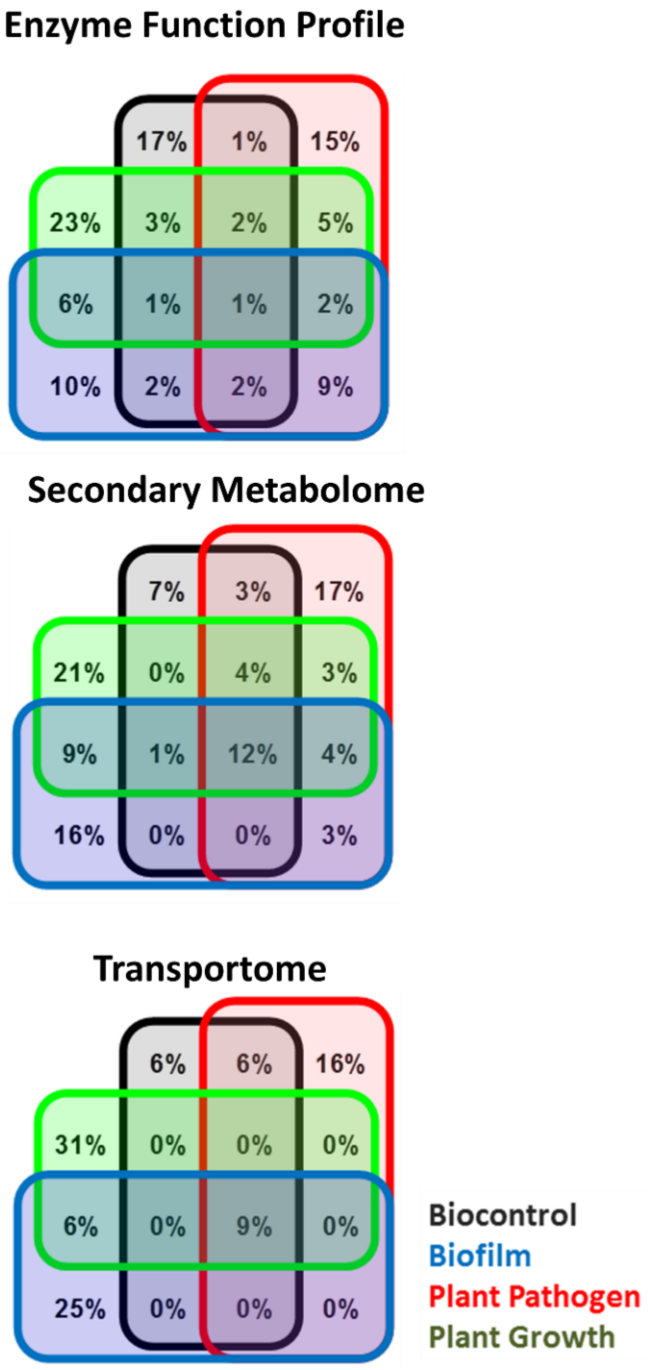


Figure 4.4. Venn diagrams for significant features identified by SVM for each data type and host interaction class for Plant-Microbe interactions. Values are presented as percent of total highly predicted features for that data type.

Strongly predictive features are not necessarily unique to a single host-interaction phenotype (**Figure 4.4** and **4.5**). While features that are unique to a specific HMI phenotype can confidently be associated with that HMI class, there are multiple reasons predictive features might belong to multiple

HMI classes. It may be that shared features comprise common molecular mechanisms for both HMI phenotypes. For example, shared predictive features between the capacities for inducing plant disease and for promoting plant growth, which could be attributed to common plant-sensing capacities or overlapping capacities in the ability to colonize plant tissues. Alternately, overlapping features may be due to co-occurrence of features in the genomes and do not derive from any biologically relevant overlap of functions. For example, if many Pseudomonads possess both Plant Growth Promotion and Biocontrol functions, then some features specific to one of the HMI classes might be misattributed to the other with some frequency.

In the following results, groupings of Plant-Microbe and Human-Microbe HMI classes will be considered separately as there is not any expected biologically relevant functional overlap between these two hosts.

4.6.1. Predicted Plant-Microbe Interaction Mechanisms

The overlap between predictive features in Plant-Microbe interaction phenotypes is summarized in a Venn diagram in **Figure 4.4**. Consideration of those features that are unique and those that are shared between more than on HMI-type provides potential insights into the different molecular mechanisms that underlie host interactions. Plant Growth Promotion has the largest proportion of uniquely predictive Enzyme Function Profile features for all data types, suggesting that this HMI phenotype is almost entirely independent of other HMI classes. Biocontrol HMI class has the least unique features by Secondary Metabolism or Transportome, but the second largest proportion of predictive Enzyme Function Profile features, suggesting that Biocontrol is a function of a few specific enzyme activities that are not well integrated into the rest of Pseudomonad metabolic pathways. Transportome has by far the least overlap between predictive HMI phenotype features, indicating that the Pseudomonad's Transportome is highly specific to its environment and its capacity to interact with a plant host.

From the observed overlap between predictive features, we propose an HMI meta-class: Survival in the Rhizosphere. The features common to all plant HMI phenotypes are hypothesized here to be those biological functions required for colonization and survival on or near plant roots, and are a prerequisite for all other plant interactions.

In **Table 4.2**, specific features associated with the plant-microbe HMI phenotypes (including Rhizosphere) are presented that can, supported by evidence in the previously published literature, provide possible specific insights into the specific molecular mechanisms that drive specific Plant-HMIs.
From this table of features, some general trends emerge. Biocontrol is not only associated with the production of compounds associated with antibacterial and antifungal characteristics, but also compounds that are proposed to directly stimulate the plant's immune system. Vitamin B12 might be a precursor for additional biosynthetic capabilities in Pseudomonad metabolism or else possibly provided directly to their plant host to induce pathogen resistance pathways. Biofilm Formation includes not just those inter-bacterial signaling compounds previously associated with biofilms, but also metabolic capacity for anti-biofilm compounds. It is possible that Pseudomonads possess this later capacity to degrade biofilm-disruption compounds synthesized by competitors in their environment, or else use them to disrupt the biofilm formation of other bacteria that compete for space in colonizing plant roots. Those features predictive for the ability to cause disease in plants are comprised of plant signaling compounds, either fatty acids or small molecules associated with plant pathogens. Plant Growth Promotion is predicted to possess two mechanisms of action: biosynthesis of plant regulatory compounds that stimulate growth directly and production of organic acids that can help to mobilize mineralized nutrients in soil, making them available for uptake by plant roots. Those features predicted to be necessary for life in this rhizosphere regulate plant stress, presumably avoiding activation plant anti-bacterial defense mechanisms, regulate osmolality and charge balance in the challenging soil environment, and take advantage of plant-specific carbon sources in the soil.

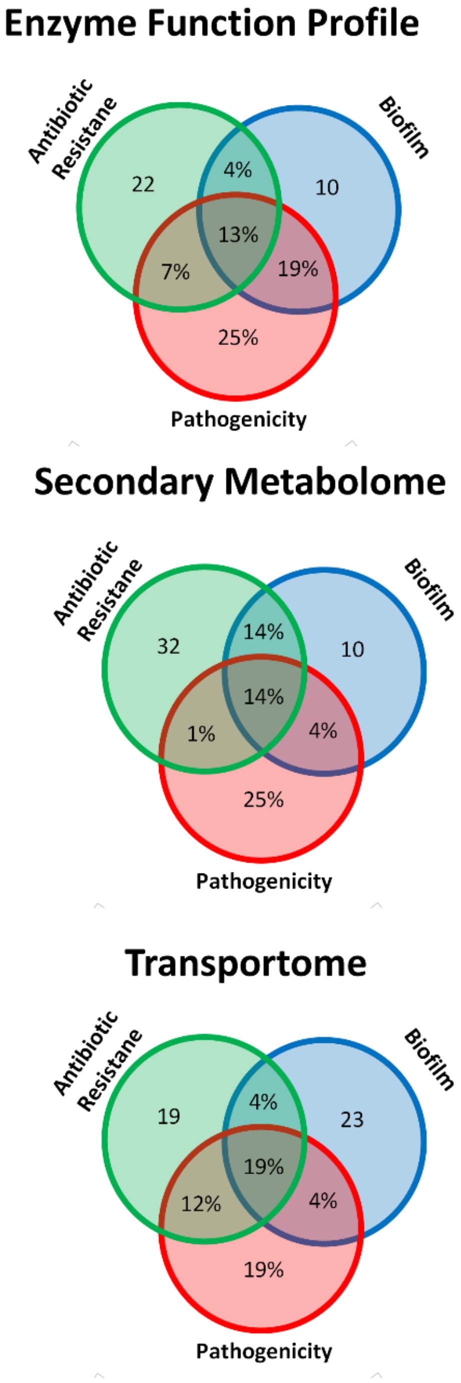


Figure 4.5. Venn diagrams for significant features identified by SVM for each data type and host interaction class for Animal-Microbe Interactions. Values are presented as percent of total highly predicted features for that data type.

4.6.2. Predicted Human-microbiome Interactions

A similar approach to analysis of the Plant-microbe interaction mechanism is applied to the Human-microbe interaction mechanisms. In spite of there being fewer HMI-phenotypes for human pathogen-related function than for plant-interactions, there is greater overlap between set of predictive features (**Figure 4.5**). When considering which features are uniquely predictive for an HMI-class and which are shares, the largest proportion of unique feature is found in Secondary Metabolism predictive for Antibiotic Resistance. This may be indicative of the range of metabolic functions requires to effectively degrade or otherwise transform antibiotics in the environment. Unique Transportomic features are roughly evenly distributed over the three HMI-phenotypes. There is a larger overlap however between Pathogen and Antibiotic Resistance, indicating that there are either several common molecular mechanisms between these HMI types or else Antibiotic resistance and Pathogenicity frequently co-occur is Pseudomonads, i.e. Pseudomonads that are pathogenic are also frequently resistant to antibiotics.

Also, similarly to the analysis of Plant-microbe predictive features, we propose a HMI metaclass, "Survival in a Human Host", which is comprised of those predictive features common to Pathogen, Antibiotic Resistance, and Biofilm Formation HMIs. If we advocate, as we did for plant-microbe interactions, that overlapping predictive features are those more general to survival in the host organism, then from these results we suggest that the human host is a more challenging environment for Pseudomonads than is the rhizosphere.

In **Table 4.3**, specific features associated with the human-microbe HMI phenotypes (including Survival in Human Host) are presented that provide insight into HMI mechanisms via mechanisms supported in previously published literature.

Pathogenicity is characterized both by specific markers of pathogenicity as well as compounds associated with reducing completion from other bacterial strains. The colicin transport function makes an appealing potential target for novel therapeutic interventions. Reducing a pathogens competitive advantage against other bacteria might permit the normal, healthy members of the microbiome to keep a pathogenic species in check without the use of non-specifically killing antibiotics. Antibiotic Resistance is dominated by efflux capacity and the metabolic capacity to degrade antibiotics. The quorum sensing features may enable colonies of bacteria to coordinate responses to antibiotics, and propose an additional mechanism to reduce antibiotic resistance. Interfering with inter-bacterial signaling may inhibit

Pseudomonads ability to coordinate responses and increase their susceptibility to antibiotics to which they may otherwise be resistant.

There are two general mechanisms proposed here for surviving within a human host. One is survival in a metal-poor environment through ion scavenging. The other is to interact directly with the host by influencing the gut-brain axis and regulating host immune responses. A potential intervention proposed by the Human Host HMI meta-class is to interfere with a pathogen's ability to gather necessary metal ions. Another is to inhibit the bacteria's capacity to influence host immune response via arginine or dopamine biosynthesis and export, potentially making the bacteria vulnerable to the host's native immune system without resorting to the use of traditional antibiotics.

5. MODELING HUMAN DYSBIOSIS

The human microbiome can have a profound effect on human health, both positive and negative. Gut microbiome communities interact with the human immune system to ward off infections, help the host to digest food, and synthesize crucial nutrients (Ramakrishna 2013, Bou Saab et al. 2014, Hennessy et al. 2014, Walsh et al. 2014). Dysbiosis occurs, however, when the microbiome has negative effects on the host's health. A dysbiotic microbiome can lead to irritable bowel syndrome, predispose its host to a variety of cancers, and increase the risks for obesity and diabetes (Collado et al. 2007, Moran and Shanahan 2014, Sanz and Moya-Perez 2014, Cox and Blaser 2015, McLean et al. 2015). An unhealthy microbiome has been implicated in depression and autism (Foster and McVey Neufeld 2013, Fond et al. 2015, Luna and Foster 2015).

Not only does the microbiome play an important part in human health, but the human microbiome is also a highly diverse community, constantly evolving in dynamic interactions with their host and the host's environment (Theriot et al. 2014, Ursell et al. 2014, Rojo et al. 2015). It is this vast diversity of the microbiome that confounds efforts to associate members of the microbiome community with effects on their host. Isolating what features are relevant to a specific HMI and what features are associated with any of the other factors that differentiate one individual from another is a substantial challenge (Fukuda and Ohno 2014, Walters et al. 2014).

Fortunately, the functional diversity of the gut microbiome is far simpler than the microbiome's taxonomic variability (Turnbaugh et al. 2009a, Consortium 2012). This suggests for a given microbiome functional capacity, there are a potentially wide range of community compositions that can provide the relevant molecular mechanisms to assemble a given HMI function. The microbiome uses a finite set of molecular tools, distributed across the genomes of its members that can be arranged and re-arranged to comprise a vast repertoire of HMI classes. Thus, HMI is less dependent upon the specific identity and relative abundance of its community members than on emergent properties of the entire microbiome.

The goal of Chapter 5 is to generate computational models of human HMIs that link microbiome community, diet, and host dysbiosis. We hypothesize that a dysbiotic HMI is an emergent property of the microbiome and that community functional profiles and metabolomes will be more predictive of specific HMIs than community structure alone. This aim will generate a system of predictive computational models that will (i) predict host dysbiosis from emergent properties of the microbiome community, (ii) predict microbiome community dynamics as a function of host diet parameters, and (iii) link the previous models in a system-sale model that will predict diet-induced dysbiosis.

5.1 Background

All of us are communities. We are complex ecosystems that have within us entire worlds of organisms that interact with us, our diet, and our environment to influence our health. Recognizing that humans are superorganisms, a single entity comprised of many, many interacting parts provides a unique insight into fighting human disease maintaining human health (**Figure 5.1**).

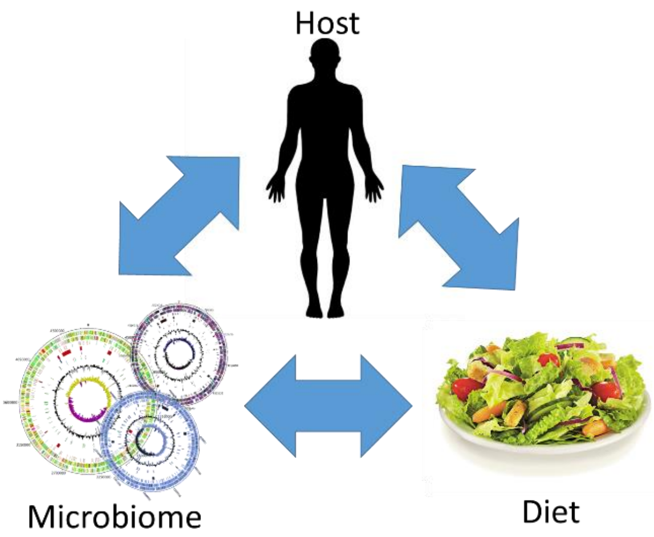


Figure 5.1. Microbiome, host, and diet interact with one another in the human superorganism. In the human ecosystem, diet and microbiome and host health are irrevocably linked into complex networks of mutual interactions. While diet can influence the microbiome community structure, the microbiome in turn affects our ability to digest and derive nutrients from food. While host behavior and environment has a profound influence on the microbiome, so too does the microbiome impact host health. Understanding and disentangling these mutual dependencies and influences between microbiome, host, and diet is a principle goal of HMI modeling.

5.1.1 Host Diet Influences Host Microbiome Community Structure

Long term changes in the host's diet can alter the microbiome's community structure. For example, the differences in the microbiome communities between meat-eating and vegetarian humans mirror the differences in the microbiomes of carnivorous and herbivorous animals (Muegge et al. 2011). These shifts in microbiome can be explained by the different metabolic requirements of protein catabolism and the fermentation of carbohydrates and indicate that the microbiome community can adapt to the changing behaviors of the host in order to optimize nutrient acquisition from the available food sources. The human microbiome can also respond to changes in the host's diet in far shorter time spans (David et al. 2014a, David et al. 2014b). Alterations in diet, in conjunction with changes in activity levels, can significantly alter the diversity and composition of the microbiome (Clarke et al. 2014). The bacteria in our changing microbiome can colonize us from a variety of sources, from within and upon the foods we eat (Economou and Gousia 2015), the environments we occupy (Lax et al. 2017), and even the other people and animals with which we associate (Lax et al. 2014).

5.1.2 Human HMI

While diet can alter host microbiome, the microbiome has a significant influence on host health. The abundance of some specific strains of bacteria is associated with a dysbiotic state. Human males with type 2 diabetes were found to have significantly lower abundances of the *Firmicutes* Clostridia and higher abundances of *Betaproteobacteria* than non-diabetic males in a different study (Larsen et al. 2010). IBS is linked to decreased levels of *Firmucutes* and *Bacteriodetes* and increased levels of *Proteobacteria* and *Actinobacteria* with an overall higher abundance of Gram-negative bacteria (Manichanh et al. 2012, Kostic et al. 2014).

5.1.3 Diet-Induced Dysbiosis

When the interactions between diet, host, and microbiome are out of balance, the result can be diet-induced dysbiosis. High-fat (Zhang et al. 2010), high-fat and high-sugar (Turnbaugh et al. 2009b), and low fiber diets (Johansson et al. 2014, Holscher et al. 2015) have all been previously identified as capable of inducing dysbiosis in a human host. The mechanism of dysbiosis has been proposed to be due, in part, to increased inflammation in the host (Brown et al. 2012a, Devkota and Chang 2013).

5.1.4 Investigating Human HMI

Numerous attempts have been made to computationally link microbiome community to specific HMI. Shotgun metagenomic data of gut microbiome populations were used to predict type 2 diabetes using a cohort of 145 64-year old European women with significant accuracy (Karlsson et al. 2013). However, when the same approach was used on a similar cohort of Chinese women, although type 2 diabetes could still be predicted with similar accuracy, the metagenomic markers between European and Chinese cohorts were very different (Qin et al. 2012). This supports the hypothesis that it is not the specific microbiome community structure itself, but rather an emergent property of the community that drives HMIs. In fact, in the conclusion of a review of microbiome and metabolic disease, specific genera, classes, or species of bacteria cannot be positively or negatively correlated with specific HMI (Fukuda and Ohno 2014). To confidently link microbiomes with HMIs, more information than microbiome community structure is needed.

5.1.5 Key Knowledge Gaps and Innovation

While previous modeling efforts have attempted to link metabolic models with microbiome community interactions and HMI mechanisms, these approaches are inherently limited and miss many of the most crucial HMI mechanisms. Not all HMI interactions can be ascribed to the metabolic consumption and transformation of host diet-derived metabolites and few bacterial interactions are limited to syntrophy or completion for biomass accumulation. Many of the HMI interaction are associated with the biosynthesis of compounds that interact directly with the host's immune, endocrine, or nervous system (Kipanyula et al. 2013, Ridlon et al. 2013, Boesjes and Brufau 2014, Dupont 2014, Carabotti et al. 2015, Gomez-Arango et al. 2016). Competition for nutrients alone cannot account for many of ways that bacteria interact in the gut, such as biofilm formation, antibiotic biosynthesis, predation, and quorum signaling (Lister et al. 2009, de Bentzmann and Plesiat 2011, Alhede et al. 2014).

Here, we apply a statistical approach combined with machine learning to computationally-derived emergent properties of microbiome communities to elucidate the molecular mechanisms of HMI. To accomplish this, we will consider community members, not as collections of specific sequenced genomes, but as a statistical distributions of genome features observed across bacterial taxonomic groupings to describe microbiome communities. We will use this novel approach to describing microbiome communities to generate superior predictions of metagenomes and apply our metabolic modeling tool, PRMT, to generate metabolic models comprised of thousands of metabolites and enzymatic transformations. The final result will be a complete, system-scale model of diet-induced dysbiosis HMI as emergent properties of the microbiome community.

5.2 Outline of Experimental Approaches

This chapter is divided into three Tasks.

Task 1: Predict dysbiosis as a function of microbiome community. “Dysbiosis” is any condition in which perturbations to a host’s microbiome leads to a negative impact on the host’s health. Microbiome community structures, however, can differ vastly from host to host and over time within the same host, making it difficult to generalize the molecular mechanisms of host-microbiome interaction that lead to dysbiosis. We propose that it is not the microbiome community structure that is most informative of host-microbiome interaction, but rather the emergent properties of the microbiome community that must be considered. A machine learning tool, e.g. SVM, can be used to predict host dysbiosis using from structure, enzyme function profile, or community metabolome data. The model also proposes the specific enzyme activities and metabolites that are predictive of host dysbiosis and provide biologically testable hypotheses for identifying specific molecular targets for therapeutic intervention.

Task 2: Generate a dynamic model of microbiome community changes in response to diet. Using our MAP-model for computational modeling of microbiome community structure as a function of environmental parameters, a significant modification to MAP-model will be created that will create dynamic models of microbiome communities. Specifically, the current host's microbiome community structure will be described as functions of the host's current diet and host's previous microbiome community structure. This dynamic model will allow us to track the evolution of human microbiome community structure over time in response to a changing host's diet.

Task 3: Predict diet-induced dysbiosis for human microbiomes. In the third task, models from Task 1 and Task 2 will be integrated into system-scale model for predicting diet-induced dysbiosis. This system-scale model will be used to predict diets that can induce dysbiosis in the host.

5.2.1 Selection of a Microbiome Dataset: A Longitudinal Study of Human Microbiome Dynamics

All Tasks in this Chapter use a common human microbiome dataset (David et al. 2014b). In a 2014 study of human microbiome communities, two donors (identified as 'Donor A' and 'Donor B' in the publication) tracked their gut microbiome community structures and recorded their diet parameters at nearly daily intervals for about one year (**Figure 5.2**). Over the course of the study, both donors coincidently experienced periods of dysbiosis. Donor A experienced dysbiosis upon international travel "to the developing world" and Donor B upon encountering food poisoning. Donor A and B microbiome populations were significantly distinct from one another, both during 'Healthy' periods and during periods of 'Dysbiosis'. Further, while upon returning to a healthy, non-dysbiosis state, Donor A's microbiome community returned to the same pre-travel community composition. Alternativrly, after dysbiosis the microbiome community of Donor B assumed a completely different microbiome community structure after dysbiosis compared to the community structure before dysbiosis. This dataset provides a truly unique opportunity for microbiome modeling. Not only are longitudinal studies of human subjects rare, but the fortuitous (for our purposes of modeling, anyway) experience of dysbiosis by both donors provides an exceptional opportunity to ask questions about the emergent properties of community that are associated with dysbiosis.

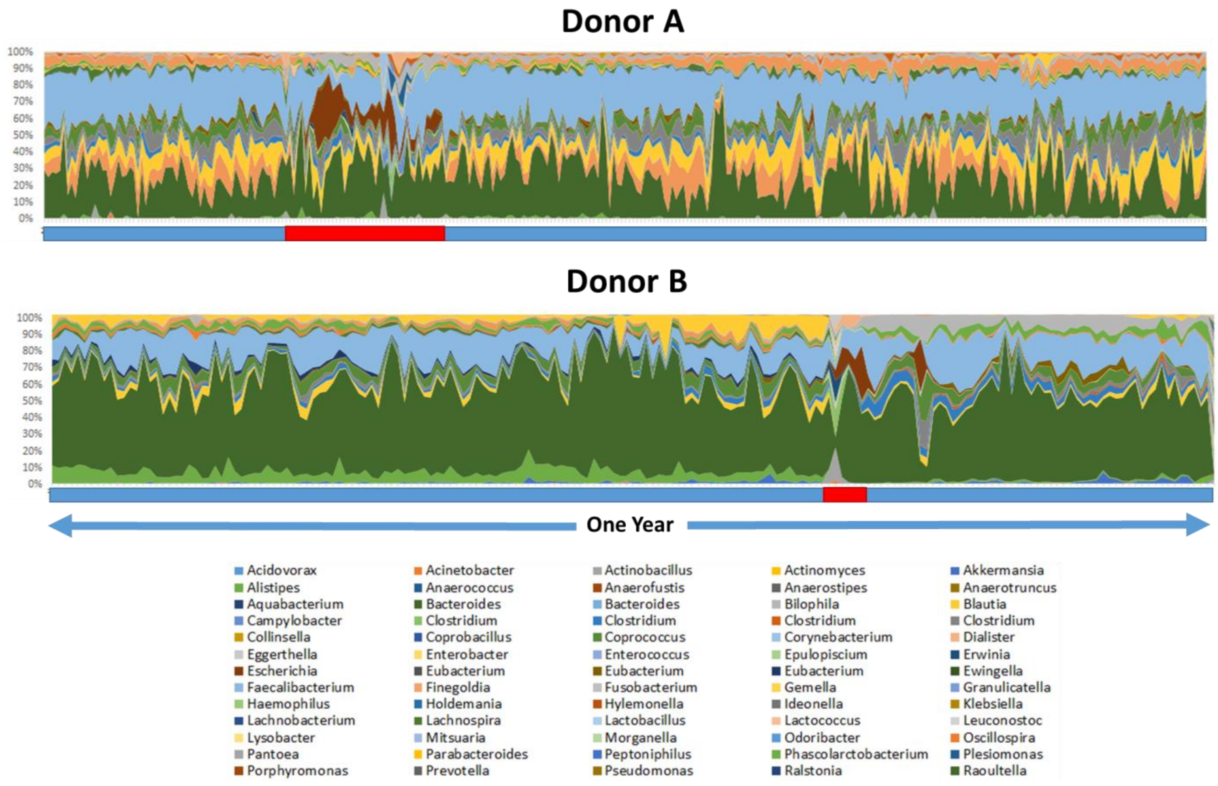


Figure 5.2. Longitudinal observations from human microbiomes. 81 bacterial genera were used to describe human microbiome communities. On x-axis, blue highlight indicated time points for which donor phenotype is 'Healthy', red highlights indicate 'Dysbiosis' microbiomes. The change in community structure during periods of Dysbiosis can be seen in these figures, as well as the shift in Healthy community structure before and after Dysbiosis in Donor B.

The observations in the dataset fit well with our prior expectations. The very different population structurers observed here precludes dysbiosis being dependent upon the presence or absence of specific bacterial species. Rather, dysbiosis HMI must be an emergent property of the community.

The dataset itself was generously provided by Dr. Lawrence David at the Duke Center for Genomics and Computational Biology.

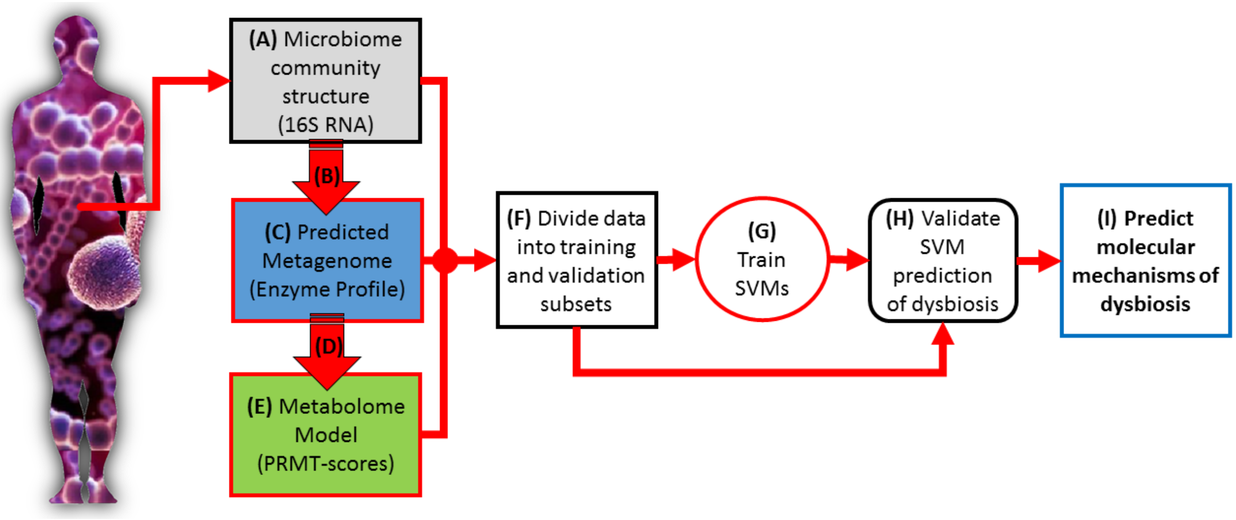

(A) Microbiome community structure (16S RNA)
(B)
(C) Predicted Metagenome (Enzyme Profile)
(D)
(E) Metabolome Model (PRMT-scores)
(F) Divide data into training and validation subsets
(G) Train SVMs
(H) Validate SVM prediction of dysbiosis
(I) Predict molecular mechanisms of dysbiosis

Figure 5.3. Outline of dysbiosis prediction approach. **(A)** Microbiome community structure data (**Figure 5.2**) is collected from the David et al. manuscript. Using the TAP-prediction approach **(B)** Enzyme Function Profiles (EFPs) are predicted from community structures **(C).** Using PRMT **(D),** microbiome community metabolomes are modeled using EFPs. Each datatype is divided into training and validation subsets **(F)**. Two methods of validation are considered by this approach: mixed donor in which training data is selected from both Donor A and B data, and cross-donor in which training data is taken from one donor and validation data is taken from the other. Training dataset is used to train SVMs for the prediction of Dysbiosis microbiomes **(G)**, and models are validated on validation subsets **(H)**. The ultimate goal of this approach is not no accurately identify Dysbiosis microbiomes, but to use the SVMs to propose the molecular mechanisms by which the microbiome induces dysbiosis in the host.

5.3 Predicting Human Dysbiosis

In the first Task, the motivation is to generate a predictive model of host dysbiosis predicted as a function of the microbiome community. The selected David et al. dataset is well suited for this purpose. The dataset contains three distinct examples of ‘Healthy’ microbiome community structures (Donor A ‘before and after travel’, Donor B ‘before illness’, and Donor B ‘after illness’) and two examples

'Dysbiosis' population structures (Donor A and Donor B). The goal of this Aim is to predict host state, Healthy or Dysbiosis, from gut microbiome data.

To propose the emergent properties of microbiome communities from community structure data, enzyme function profile data and metabolomic model data were generated. Microbiome data was used to train SVM models to predict dysbiosis. It is hypothesized that community structure should be a weak predictor of dysbiosis and that emergent properties of microbiome community such as microbiome community metabolome will be a better predictor of dysbiosis. The overall approach taken by this Task is summarized in **Figure 5.3**. Many of the results of this analysis have been previously published in (Larsen and Dai 2015).

5.3.1 Data selection and Prediction of Microbiome Emergent Properties

From the David et al. longitudinal survey of human microbiome analysis (David et al. 2014a), a subset of microbiomes was selected. From these selected microbiomes, the emergent properties of the microbiome communities, the EFP and metabolome, were calculated.

5.3.1.1 Describe the Selected Microbiome Dataset at the Taxonomic level of Genera

As has been done in previous microbiome analysis methods (Larsen et al. 2012a, Larsen and Dai 2015, Larsen et al. 2015b), community structure was described at a higher taxonomic level than either species or OTU. When considered at the level of OTU, the observed population structures are very sparse for many bacteria, with many OTUs only infrequently observed, appearing in just a few observations. For this analysis, the taxonomic level of Genera was chosen for describing the population structure. There were 442 to bacterial Genera identified in the initial data across all microbiomes. For use in subsequent modeling steps, only the top-most 81 abundant Genera, selected to account for more than 99.5% of all population abundances, with the remaining bacteria comprising the final 0.05% abundance of the microbiome communities not incorporated into the models. Populations community structures were normalized such that total population abundances always sum to 100.

Due to the significant imbalance between Healthy and the much less prevalent Dysbiosis microbiome, for modeling analysis a subset of the data was considered that represents of more balanced distribution between Healthy and Dysbiosis microbiomes (**Table 5.1**). Microbiomes were selected from each category at random except for Donor B Dysbiosis. All 7 Dysbiosis Donor B microbiomes were used for analysis.

Table 5.1. Microbiome class distribution for prediction of dysbiosis

Microbiome Class	**Donor A**	**Donor B**
"Healthy" (Before Dysbiosis)	15	15
"Dysbiosis"	13	7
"Healthy" (After Dysbiosis)	15	15

5.3.1.2 Generate Enzyme Function Profiles

Using the approach described in Chapter 3, selected microbiome community structure data was used to generate Enzyme Function Profiles (EFPs). Taxonomic Average Profile (TAP) matrix was generated from available database of 2888 sequences and annotated genomes and 2055 unique enzyme functions collected from the PiCRUST database (Langille et al. 2013). Of the 81 genera in microbiome community structures, 15 did not have representation at the level of genera in the set of genomes so these bacteria were considered at the level of order for the purpose of EFP predictions. In total, there were 2340 sequenced genomes with taxonomies represented in the selected microbiome dataset.

5.3.1.3 Predict Microbiome Community Metabolomes

The resulting EFPs utilized the average genomic information of the selected 81 genera to EFPs were used to generate metabolic models using the PRMT method, as described in Chapter 2. A Secondary Metabolome was also generated with PRMT using only enzymatic reactions found in KEGG

Pathway 01110, “Biosynthesis of Secondary Metabolites” (http://www.genome.jp/kegg-bin/show_pathway?map01110). The final metabolomic model for Donor A and B microbiome metabolome has 2824 metabolites connected by 4284 enzymatic transformations mediated by 1897 unique enzyme functions (**Table 5.2**).

Table 5.2. Four datatypes used for predicting dysbiosis from microbiome

Data Type	**Number of Features**
Community Structure	81
Enzyme Function Profile	2,055
Metabolism	1,492
Secondary Metabolism	122

5.3.1.4 Comparison of Collected Microbiome Datatypes

Analysis to this point has yielded four datatypes that can be used to train models for prediction of Dysbiosis from gut microbiome: Community Structure, EFP, Metabolism, and Secondary Metabolism.

To visualize the relative degree of similarity between individual microbiomes and to determine if Dysbiotic microbiomes are distinct from Healthy microbiomes, Multidimensional Scaling Plots (MDS) were generated. MSD Plots were generated in R and results are shown in **Figure 5.4**.

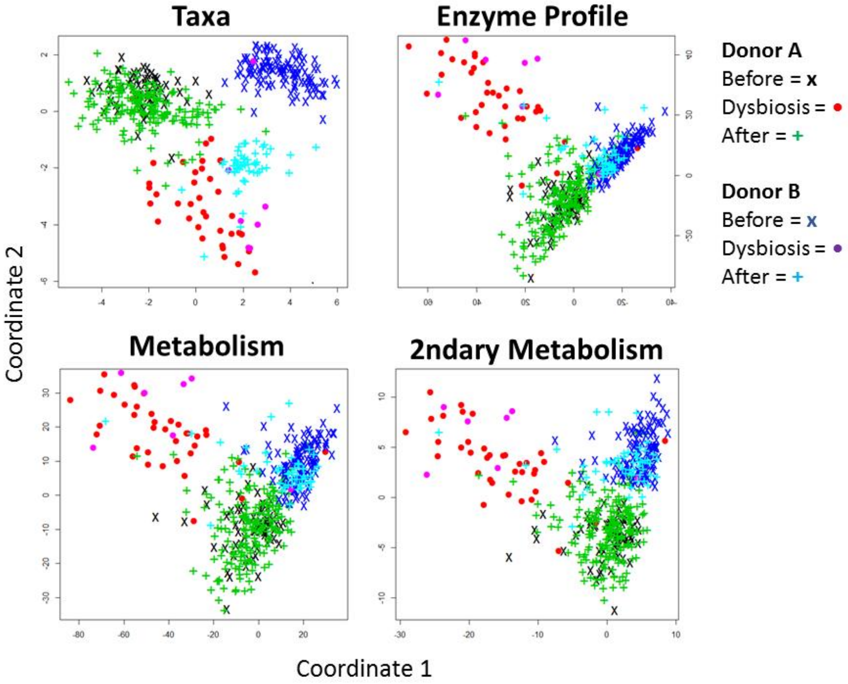


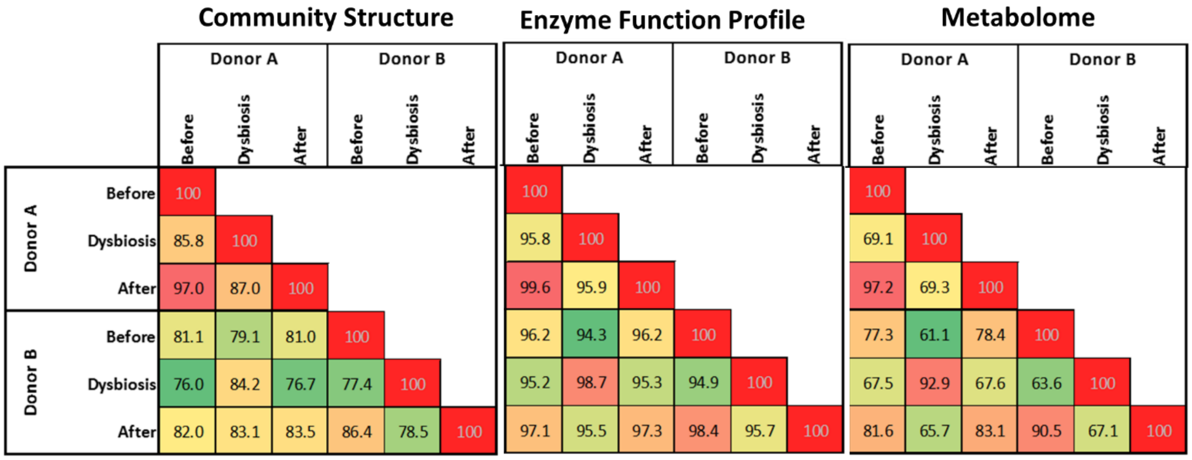
Figure 5.4. MDS Plots of human microbiome data types. Each point in an MDS plot is a single observation. Points are identified as 'Before Dysbiosis', 'Dysbiosis', and 'After Dysbiosis' and identified as being from either Donor A or Donor B.

Visualization of data by MDS plots supports the expectation that dysbiosis is an emergent property of the microbiome. When plotted by Genera, Healthy and Dysbiosis microbiomes do not separate well and the best grouping is by Donor. For Enzyme Function Profile data, Healthy and Dysbiosis microbiomes more clearly form separate groups. Metabolism and Secondary metabolism show the best separation between Dysbiosis and Healthy microbiome microbiomes with the best separation observed for the complete Metabolism datatype.

A more quantitative approach for comparing datatypes was also considered. The Bray-Curtis (BC) similarity between the average microbiome values for each datatype was calculated. BC dissimilarity score is a statistical measurement from ecology that compares the relative species compositions from two locations. The BC dissimilarity is calculated as:

$$BC = \frac{\sum_{i=1}^{species} |site1_i - site2_i|}{\sum_{i=1}^{species} site1_i + site2_i} \qquad \textbf{Eq. 5.1}$$

Where *species* is the number of species/counted taxa in the populations being compared and *site1*$_i$ and *site2*$_i$ are the abundances of the i^{th} taxa at site1 and site2 respectively.



Community Structure

		Donor A Before	Donor A Dysbiosis	Donor A After	Donor B Before	Donor B Dysbiosis	Donor B After
Donor A	Before	100					
	Dysbiosis	85.8	100				
	After	97.0	87.0	100			
Donor B	Before	81.1	79.1	81.0	100		
	Dysbiosis	76.0	84.2	76.7	77.4	100	
	After	82.0	83.1	83.5	86.4	78.5	100

Enzyme Function Profile

		Donor A Before	Donor A Dysbiosis	Donor A After	Donor B Before	Donor B Dysbiosis	Donor B After
Donor A	Before	100					
	Dysbiosis	95.8	100				
	After	99.6	95.9	100			
Donor B	Before	96.2	94.3	96.2	100		
	Dysbiosis	95.2	98.7	95.3	94.9	100	
	After	97.1	95.5	97.3	98.4	95.7	100

Metabolome

		Donor A Before	Donor A Dysbiosis	Donor A After	Donor B Before	Donor B Dysbiosis	Donor B After
Donor A	Before	100					
	Dysbiosis	69.1	100				
	After	97.2	69.3	100			
Donor B	Before	77.3	61.1	78.4	100		
	Dysbiosis	67.5	92.9	67.6	63.6	100	
	After	81.6	65.7	83.1	90.5	67.1	100

Figure 5.5. Bray-Curtis dissimilarity between the average microbiome community datatypes. A BC score of 100 indicates identical populations and a score of 0 indicated that there are no common elements between populations. Values in figure a highlighted such that highest values are red and lowest values are green.

Comparing microbiome datatypes using BC dissimilarity allows us to quantify many of the observations that could be inferred qualitatively from MDS plots (**Figure 5.5**). Considering microbiome populations by their community structure, it is seen that before and after dysbiosis population for Donor A are very similar while for donor B, populations after dysbiosis are distinct. By community structure, Donor A and Donor B Healthy populations are about as different from one another as are Healthy from

Dysbiosis populations. By EFP, the populations become much more similar, supporting the hypothesis those populations very different by their population composition may still be highly similar by their functional compositions. Most notable here, it can be seen that while Donor B's before and after Dysbiosis population were distinct by taxonomic composition, they are very similar in their functional composition. While microbiome populations were found to increase similarity by EFP relative to similarity by population structure, microbiome metabolome finds increased diversity between populations. This indicates that the small differences in populations by EFP can translate into quite significant differences in metabolomic capacities. By metabolome, we find that while Healthy microbiomes are somewhat divergent between donors, Dysbiosis microbiomes are notably more similar in their metabolomic capacity. This result suggests that while there are many possible metabolomic functional compositions that can yield a Healthy microbiome, Dysbiosis microbiomes are quite similar to one another in their misery. This result is particularly striking given that the events inciting dysbiosis in Donors were considerably different, i.e. international travel vs. food poisoning.

Now that it has been established that there are differences between Dysbiosis and Healthy microbiomes, and the nature and degree of those differences is dependent upon the microbiome datatype considered, the next step in analysis is to identify which microbiome datatype features are most predictive for distinguishing Healthy vs. Dysbiosis microbiomes.

5.3.2 Select Microbiome Features Most Predictive of Dysbiosis

Here, the goal is to identify the microbiome features most predictive for dysbiosis. The tool used for prediction in this section is SVM.

For each of the four datatypes, features were ranked by Fisher-score. Fisher-scores, for a given datatype feature *i*, are calculated as:

$$Fisher_i = \frac{|\mathbf{AVERAGE}(Healthy_i) - \mathbf{AVERAGE}(Dysbiosis_i)|}{\mathbf{STDEV}(AllFeatures_i)} \qquad \textbf{Eq. 5.2}$$

Where *Fisher*$_i$ is the Fisher score for feature i, AVERAGE(*Healthy*$_i$) is the average value of all feature *i* from Healthy microbiomes, **AVERAGE**(*Dysbiosis*$_i$) is the average value of all feature *i* from all Dysbiosis microbiomes, and **STDEV**(*AllFeatures*$_i$) is the standard deviation of feature *i* across all observations.

SVMs were used to predict Dysbiosis class microbiomes. A separate SVM was constructed for each datatype, Genera, EFP, Metabolism, and Secondary Metabolism. As described initially in Chapter 4, package ‘e1071’ v1.6-1 in R-project was used to make SVMs. SVMs used linear kernels and 10-fold cross validation. SVMs were trained using the 60 Healthy and 20 Dysbiosis microbiomes (**Table 5.1**). Models were validated on the remaining microbiome data, comprised of 375 microbiomes of class ‘Healthy’ and 22 microbiomes of class ‘Dysbiosis’.

Subsets of datatypes were used to identify most predictive features from each datatype. SVMs were trained on the top most 100, 90, 80, 70, 60, 50, 50, 30, 20, and 10% of features as ranked by Fisher score. For datatypes EFP and Metabolism, additional subsets from the top 5, 2.5, 1.25, and 0.625% Fisher-score ranked features were also considered. Results are reported as F-scores for each subset of each datatype (**Figure 5.6**).

While all datatypes and subset sizes produced fairly good predictive results (F-scores > 0.75 in all cases), the most predictive subset for each datatype demonstrated very good predictive power for identifying Dysbiosis-class microbiomes from every datatype. The most predictive subset sizes are listed in **Table 5.3**.

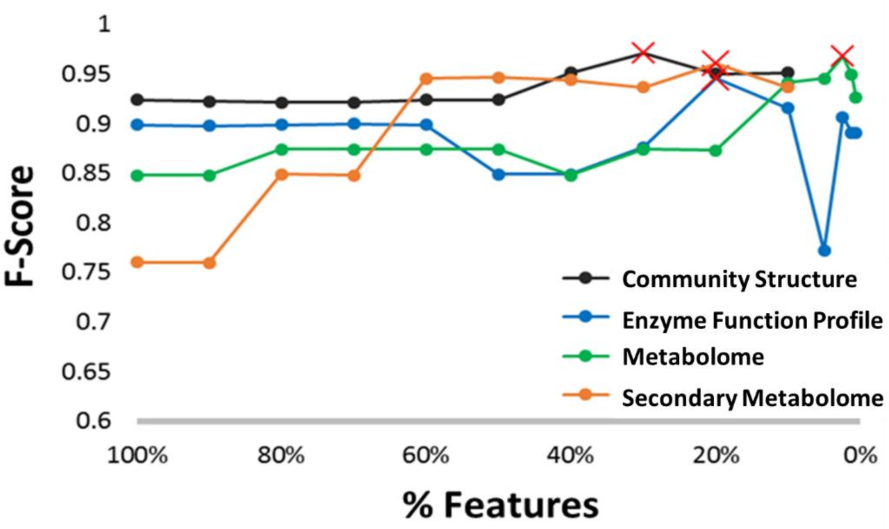


Figure 5.6. Identifying the most predictive features for dysbiosis from multiple microbiome datatypes. Four microbiome datatypes, Genera, EFP, Metabolism, and Secondary Metabolism, were ranked by Fisher score and SVM were trained on differently sized sets to find subset that best predicts Dysbisois microbiomes, where prediction is ranked by F-score. In figure, red 'X's indicate the datatype subset that gave the highest F-score for validation microbiomes. Values and datatype subset sizes for best predictions are listed in **Table 5.2**.

Table 5.3. Number of features in subsets most predictive for Dysbiosis

Data Type	# Features	F-score
Community Structure (Genera)	24	0.97
EFP	380	0.95
Metabolism	36	0.97
Secondary Metabolism	24	0.96

Contrary to what might have been expected from MDS plots and BC-dissimilarity scores, there is no difference in the predictive power of metabolome over genera for identifying Dysbiosis microbiomes. Based upon MDS and BC results, it may have been expected that Metabolome would be more predictive that Genera. While the identified datatype subsets will be analyzed in subsequent sections to propose possible molecular mechanisms of dysbiosis, first the relative predictive power of different microbiome datatypes will be considered using a cross-donor validation approach.

5.3.3 Cross-Donor Validation of Dysbiosis Prediction

We have previously proposed that dysbiosis in the human microbiome, and HMIs in general, are emergent properties of the microbiome population. According to this hypothesis, it is not the abundance or absence of a particular bacterium that drives HMIs, but rather the results of the entire community and the particular set of biological functions that is manifest by the community as a whole. The previous results for predicting dysbiosis from multiple microbiome datatypes appears not to support this hypothesis; community composition was no more and no less predictive that the emergent property of microbiome metabolome. Here, we delve deeper into this observation by considering a new validation scheme for dysbiosis prediction: Cross-donor validation.

In the cross-donor validation scheme, SVM predictors are trained on data from one donor and validated on data from the other donor. For training data, the most predictive features of each datatype identified in the previous section will be used. The same SVM parameters will be used as described in the previous section and predictive power of results are expressed as F-scores. Results are summarized in **Figure 5.7**.

Results of cross-donor validation show a range of predictive outcomes. SVM trained on Genera are the least predictive with SVM trained on Donor B and validated on Donor A fairing particularly poorly. This makes intuitive sense as Donor B had the most varied Healthy microbiomes, with before and after dysbiosis populations being very different from one another. The emergent properties of EFP, Metabolome and Secondary Metabolome perform much better in the cross-donor validation approach. In spite of the significant differences in observed population structure between donors, these datatypes show very strong abilities to predict dysbiosis in the alternate donor. Though the absolute predictive abilities varies by training set and by datatype, from these results Metabolome is the most predictive datatype, followed by Secondary Metabolome and then Enzyme Function Profile.

Contrary to the mixed donor data results in the previous section, here we find our main hypothesis well supported: dysbiosis and HMI are emergent properties of microbiome communities.

Population structure alone has only a poor capacity to predict an HMI type when faced with the variety of human microbiome communities.

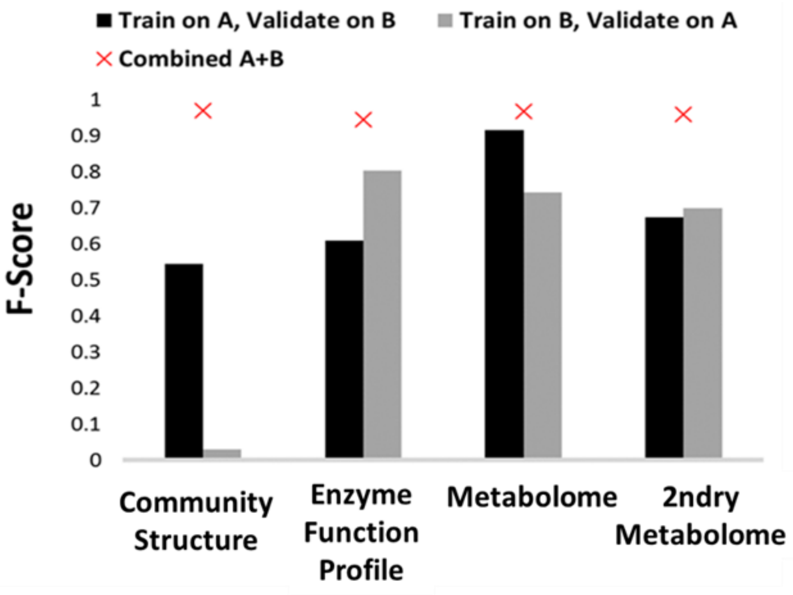


Figure 5.7. Cross-donor dysbiosis prediction results. Red 'X's indicate the F-score for combined donor results, shown in Figure X and reproduced here to show cross-donor validation results in context of results for mixed-donor validation (from Figure X).

5.3.4 Enriched Metabolic Pathways in Dysbiotic Microbiomes

While it has now been demonstrated that the emergent properties of microbiome communities can be effective predictors of dysbiosis, the utility of this approach as a diagnostic is of questionable value. If the goal of an analysis is to identify a state of dysbiosis in human subjects, then one might imagine simpler observations that detailed molecular characterization of their microbiomes. The more relevant value of this approach is in the identification of the predictive microbiome features of dysbiosis, for by using those features, the molecular mechanisms by with the microbiome can induce dysbiosis upon its host may be deduced.

While the complete list of specific predictive features for dysbiosis can be found in (Larsen and Dai 2015), simply listing those features does not provide much insight into the potential set of biological functions they represent. To place the predictive features into the broader biological context, the specific KEGG pathways enriched for sets of dysbiosis predictive features, relative to the total features in each datatype was calculated. Enrichment for a KEGG pathway for a given datatype was calculated as a cumulative Hypergeometric Distribution:

$$Enrichment_P = 1 - \sum_{i=1}^{n} \frac{\binom{K}{k}\binom{N-K}{i-k}}{\binom{N}{i}} \qquad \textbf{Eq. 5.3}$$

Where *Enrichment*$_P$ is the p-value for enrichment of KEGG Pathway *P*, *N* is the total number of microbiome features, n is the number of predictive microbiome features, *K* is the total number of microbiome features that map to KEGG pathway *P*, and *k* is the number of predictive features that map to KEGG pathway *P*. An enrichment value close to 0 indicates a significant enrichment. A threshold of p-value less than 0.05 was used to determine significance of KEGG pathway enrichment. Tables of predictive features and enriched KEGG Pathways follow for Enzyme Function Profiles (**Table 5.4**), Metabolome (**Table 5.5**), and Secondary Metabolome (**Table 5.6**).

Table 5.4. Enriched pathways in most predictive community enzyme function profile features

KEGG ID	Pathway	Enzyme Functions	Enrichment p-Val
map00121	Secondary bile acid biosynthesis	1.-.-.-, 4.2.1.-, 6.-.-.-	0.00E+00
map01053	Biosynthesis of siderophore group nonribosomal peptides	1.3.1.28 , 3.3.2.1, 2.7.7.58, 6.3.2.-	1.08E-02
map00540	Lipopolysaccharide biosynthesis	2.4.1.56, 2.4.-.-, 3.6.1.- , 2.7.1.-, 5.1.3.20, 3.1.3.-, 5.-.-.-, 2.3.1.-, 6.-.-.-, 2.4.1.44	1.65E-02
map00904	Diterpenoid biosynthesis	1.14.11.-, 1.14.13.-, 2.3.1.-	3.06E-02
map00053	Ascorbate and aldarate metabolism	4.2.1.42, 4.1.1.85, 4.2.1.40, 3.1.1.-, 4.1.2.20, 3.7.1.-, 5.1.3.22, 1.1.1.122, 1.1.1.130, 3.1.3.-, 5.1.3.4, 2.7.1.53	3.45E-02
map00480	Glutathione metabolism	3.5.1.78, 3.4.11.23, 4.1.1.17, 6.3.2.3, 1.8.1.7, 1.17.4.1, 2.5.1.18, 2.3.2.2, 6.3.1.8 , 3.5.2.9	3.76E-02
map00906	Carotenoid biosynthesis	1.-.-.-, 2.5.1.-, 1.14.13.-, 5.-.-.-, 2.3.1.-	4.68E-02

Table 5.5. Enriched pathways in most predictive total community metabolome model features

KEGG ID	Pathway	Metabolites	Enrichment p-Val
map00770	Pantothenate and CoA biosynthesis	CoA, Pantetheine 4'-phosphate, Apo- acyl-carrier-protein	2.86E-04
map00561	Glycerolipid metabolism	Phosphatidate, Diglucosyl-diacylglycerol, Glycerophosphoglycoglycerolipid	5.16E-04
map00030	Pentose phosphate pathway	5-Phospho-alpha-D-ribose 1-diphosphate, D-Ribose 1,5-bisphosphate, 2-Dehydro-3-deoxy-6-phospho-D-gluconate	6.71E-04
map00361	Chlorocyclohexane and chlorobenzene degradation	2-Maleylacetate, 2,4-Dichlorophenol, cis-2-Chloro-4-carboxymethylenebut-2-en-1,4-olide, 2-Chloromaleylacetate	2.57E-03
map00240	Pyrimidine metabolism	, 5-Phospho-alpha-D-ribose 1-diphosphate, Thymine	4.72E-03
map00362	Benzoate degradation	2,3-Dihydroxybenzoate, S-Benzoate coenzyme A, 2-Maleylacetate	6.56E-03
map00627	Aminobenzoate degradation	2,3-Dihydroxybenzoate, S-Benzoate coenzyme A, 2-Maleylacetate	6.56E-03
map01120	Microbial metabolism in diverse environments	5-Phospho-alpha-D-ribose 1-diphosphate, 2,3-Dihydroxybenzoate, S-Benzoate coenzyme A, 2-Maleylacetate, 2,4-Dichlorophenol, 5,10-Methenyltetrahydromethanopterin, 5,10-Methylenetetrahydromethanopterin, 2-Dehydro-3-deoxy-6-phospho-D-gluconate, cis-2-Chloro-4-carboxymethylenebut-2-en-1,4-olide, Aerobactin, Ectoine, 2-Chloromaleylacetate, 2-Hydroxy-cis-hex-2,4-dienoate, 4-Fluoromuconolactone, 2-Chloro-5-methylmaleylacetate	1.57E-02

Table 5.6. Enriched pathways in most predictive secondary community metabolome model features

KEGG ID	Pathway	Secondary Metabolites	Enrichment p-Val
map01061	Biosynthesis of phenylpropanoids	L-Tryptophan, p-Coumaroyl-CoA, Coniferyl alcohol, 4-Coumarate, Caffeate, Ferulate, Coniferyl aldehyde, 4-Hydroxycinnamyl aldehyde, 5-Hydroxyferulate, 5-Hydroxyconiferaldehyde	7.93E-07
map01120	Microbial metabolism in diverse environments	5-Phospho-alpha-D-ribose 1-diphosphate, 2,3-Dihydroxybenzoate, S-Benzoate coenzyme A, 2-Maleylacetate, 2,4-Dichlorophenol, 5,10-Methenyltetrahydromethanopterin, 5,10-Methylenetetrahydromethanopterin, 2-Dehydro-3-deoxy-6-phospho-D-gluconate, cis-2-Chloro-4-carboxymethylenebut-2-en-1,4-olide, Aerobactin, Ectoine, 2-Chloromaleylacetate, 2-Hydroxy-cis-hex-2,4-dienoate, 4-Fluoromuconolactone, 2-Chloro-5-methylmaleylacetate	1.57E-02
map00940	Phenylpropanoid biosynthesis	p-Coumaroyl-CoA, Coniferyl alcohol, 4-Coumarate, Caffeate, Ferulate, Coniferyl aldehyde, 4-Hydroxycinnamyl aldehyde, 5-Hydroxyferulate, 5-Hydroxyconiferaldehyde, 5-Hydroxyconiferyl alcohol, N1,N5,N10-Tri- hydroxyferuloyl -spermidine	1.47E-06
map04974	Protein digestion and absorption	L-Tryptophan, L-Leucine, Tyramine	1.26E-02

Interestingly, although each of the datatypes for Enzyme Function Profile, Metabolome, and Secondary Metabolome have very similar predictive abilities by F-score (**Figure 5.7**), each datatype proposes an entirely different set of enriched pathways. When combined, the three datatypes generate a more system-scale picture of the molecular mechanisms of host dysbiosis, proposing several putative molecular mechanisms by which the microbiome induces dysbiosis in its host.

5.3.5 Predict Molecular Mechanisms of Dysbiosis

From the sets of enriched pathways in the lists of features predictive for dysbiosis, it is possible to draw a number of possible hypotheses for the molecular mechanisms of dysbiosis.

5.3.5.1 Vitamin Metabolism is Altered in the Dysbiotic Metabolome

Disruption is the biosynthesis of vitamins that are important to the host is one of the mechanism proposed by which host dysbiosis occurs (Ursell et al. 2014, Yoon et al. 2015). Vitamin-associated KEGG Pathways for "Pantothenate and Co biosynthesis" (vitamin B) (**Table 5.5**), "Ascorbate and aldarate metabolism" (vitamin C) (**Table 5.4**), and "Carotenoid biosynthesis" (antioxidants) (**Table 5.4**) are all enriched in the dysbiotic microbiome.

5.3.5.2 Dysbiosis Affects Host's Digestion

One mechanism of dysbiosis that appears in these results is a disruption in the host's ability to digest or extract nutrients from food. Enriched KEGG pathways "Biosynthesis of phenylpropanoids", "Phenylpropanoid biosynthesis" (Russell et al. 2013) (**Table 5.6**), and "Protein digestion and absorption" (**Table 5.6**), and metabolites putrescine and spermidine (Table 4) implicate changes in the capacity to digest proteins in dysbiosis (Larque et al. 2007). Likewise, enrichment for the pathways "Glycerolipid metabolism" (**Table 5.5)** and "Secondary bile acid biosynthesis" (**Table 5.4**) indicate that the capacity to digest fatty acids is affected. Secondary bile acids is a particularly provoking observation as are secondary bile acids are those that result directly from bacterial metabolic activities on primary, host-synthesized bile acids.

5.3.5.3 Dysbiosis is Associated for Markers of Bacterial Virulence

The enriched KEGG pathway for "Biosynthesis of siderophore group nonribosomal peptides" (**Table 5.4**) suggests the importance of bacterial virulence factors in causing dysbiosis (Oves-Costales et al. 2009, Garenaux et al. 2011). The enriched KEGG pathways "Aminobenzoate degradation "and "Benzoate degradation" (**Table 5.5**) are implicated in IBS (Rossi et al. 2011, Rooks et al. 2014).

5.3.6 Summary of Results

Linking HMIs to microbiome community structures is confounded by the native variability of the human microbiome. Here, we show that through emergent properties of the microbiome derived from metagenome predictions and metabolome modeling, the HMI type Dysbiosis can be effectively predicted from microbiome community structure. More importantly, these predictions provide insight into the putative mechanisms of dysbiosis, potentially pointed the way to microbiome-based therapies to protect patients from the consequences of dysbiosis. Disruption of vitamin biosynthesis, protein and fatty acid digestion, and known virulence have been identified here as potential molecular mechanisms of dysbiosis in the human microbiome. These predictions are well supported by previously publications, lending credibility to these results.

Predictions of human dysbiosis, however, is the result of the analysis of only two individuals. While this is a necessary consequence of the limited availability of human longitudinal microbiome data, it is not clear that these results can be generalized beyond the individuals in this experimental dataset. Directly addressing these concerns, Chapter 6 utilizes microbiome data collected laboratory model of HMI that enables microbiome studies with far greater biological replication.

5.4 Model a Dynamic Microbiome Population as Function of the Host's Diet

The gut microbiome changes in response to host diet, identifying a potential method that can be used to deliberately manipulate the microbiome into a desired community structure (Varankovich et al. 2015, Cockburn and Koropatkin 2016) . Here, we used the David et al. microbiome data (David et al. 2014a) to generate a dynamic computational model of gut microbiome community structure as a function of host diet.

An opportunity to make a significant improvement in the MAP-model for predicting microbiome community structures as a function of environmental parameters is exploited in the current analysis. In previous application of the MAP-models (Larsen et al. 2012a, Larsen et al. 2015b), the predicted microbiome community structure was a function of that time point's environmental parameters. This is appropriate in marine environments, where the current continuously renews the environment or in soil microbiomes that are sampled only infrequently. The longitudinal nature of the available David et al. microbiome data however makes it possible to create a dynamic MAP-model, which is significantly distinct from previously published models. In the dynamic MAP-model, all bacterial abundances are calculated as a function of diet parameters at the current time pont and the population structure at the prior time point. This creates a predictive model of microbiome community structure that evolves over time as the previous population structure is able to inform the population structure of the subsequent time point in response to a changing host diet.

In this Task, there are three principle steps: (i) define longitudinal human microbiome community structure at level of Order, (ii) generate a CIN from microbiome community structure and host's diet, and (iii) use the CIN as the scaffold for a dynamic MAP-model of human microbiome community in response to changes in the host's diet.

5.4.1 Microbiome Data Selection and Data Pre-Processing

Again, we turn to the David et al. microbiome dataset (David et al. 2014a) as a valuable source of longitudinal microbiome community data. For this analysis, we restrict ourselves to the provided Donor A microbiome and diet parameters. Donor B is less appropriate for this analysis for several reasons: there are far few available data points for Donor B for which both microbiome community and diet parameters are available and the significant discontinuity in Donor B's microbiome community structure after Dysbiosis period makes diet manifestly a less important contributor to community structure dynamics in this dataset.

There are 140 Healthy-state microbiomes for Donor A that have accompanying diet parameters. For this model, we considered the population structure at the taxonomic level of Order. Of the 68 taxonomic orders detected in the dataset, only the 20 most abundant Orders, accounting for over 99.5% of the bacterial population, are used in the model (**Table 5.7**). The 48 Orders comprising the remaining 0.05% of population abundance were discarded for the remainder of this analysis. Population abundances for each observation were normalized to sum to 100 and values were $\log_2$ transformed (**Figure 5.8**). To describe host diet, there are 10 available parameters: calcium, carbohydrates, cholesterol, fat, fiber, protein, saturated fats, sodium, and sugar (**Table 5.7**). Units of diet parameters were not provided with lists of nutrient parameters and all measurements were normalized to arbitrary values between 20 and 80 (**Figure 5.8**).

Table 5.7. Taxonomic identification of microbiome community structure

Order	**Number of Genomes**	**Brief Description**
Actinomycetales	191	*Actinomycetales* are gram positive bacteria with complex cell wall structures. Many species of *Actinomycetes* produce antimicrobial compounds.
Bacteroidales	81	Uncultured bacteria found in animals.
Bifidobacteriales	40	*Bifidobacterium* are gram-positive, non-motile, anaerobes found in the human gut, vagina, and oral cavity. *Bifidobacteria* are sometimes used as probiotics.
Burkholderiales	118	*Burkholderiales* is an Order of Gram-negative bacteria that includes several pathogens.
Campylobacterales	81	The *Campylobacterales* are gram-negative microaerophiles.
Caulobacterales	10	*Caulobacteraceae* are gram-negative bacteria.
Clostridiales	166	*Clostridiales* are Gram-positive bacteria commonly found in healthy guts.

Coriobacteriales	15	*Coriobacteriales* commonly found in healthy gut microbiomes and are present at low abundance in IBS.
Desulfovibrionales	14	The *Desulfovibrionales* are Gram-negative obligate anaerobes. The majority of .*Desulfovibrionales* reduce sulfur.
Enterobacteriales	234	The *Enterobacteriales* are Gram-negative facultative anaerobes. It is frequently found on human skin, where it is 400% more abundant in females than males.
Erysipelotrichales	10	The *Erysipelotrichia* are commonly found in the human gut microbiome and increase in abundance with a high-fat diet.
Fusobacteriales	18	*Fusobacteriales* are obligate anaerobes found in in the microbiomes for human gut, lungs, mouth, and urinary tract. They are in low abundance in the gut microbiome in Crohn’s Disease.
Lactobacillales	307	*Lactobacillales* are Gram-positive, acid-tolerant bacteria that produce lactic acid as the major metabolic end product of carbohydrate fermentation.
MIZ46	0	An uncharacterized bacteria.
Pasteurellales	52	*Pasteurellales* are Gram-negative bacteria commonly found in the gut. A few can be pathogens.
Pseudomonadales	65	The *Pseudomonadales* are common in soils and a few are opportunistic pathogens.
Rhizobiales	105	The Rhizobiales are an order of Gram-negative *Alphaproteobacteria.*
Rhodospirillales	23	The *Rhodospirillales* produce acetic acid during respiration.
Erysipelotrichales	10	*Erysipelotrichia* are common in the gut microbiome, and are found at higher abundance in response to a high-fat diet.
Xanthomonadales	26	The *Xanthomonadales* are gram negative obligate aerobes.

Bacterial information collected from the NCBI Taxonomy Browser (https://www.ncbi.nlm.nih.gov/taxonomy)

Table 5.8. Diet parameters is Human longitudinal microbiome data

Human Diet Parameters
Calcium
Calorie
Carbohydrate
Cholesterol
Fat
Fiber
Protein
Saturated Fat
Sodium
Sugar

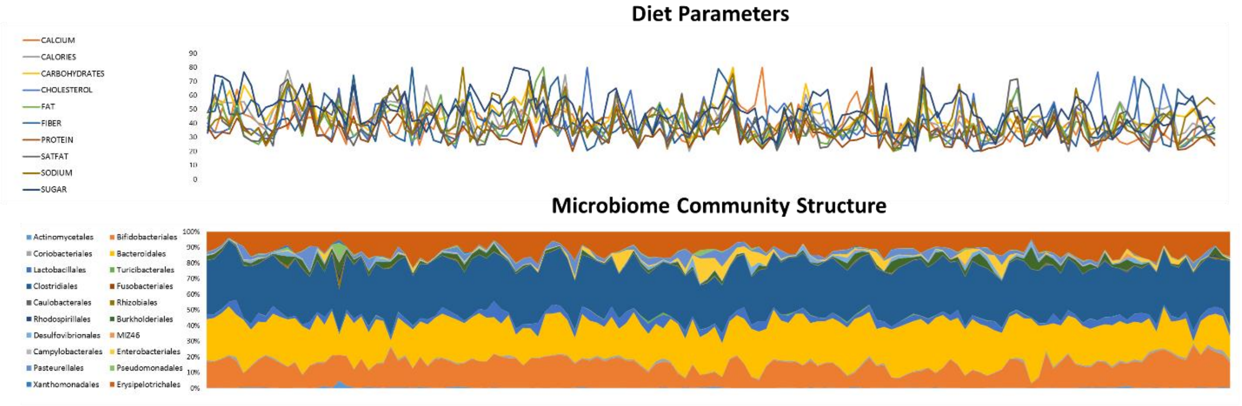


Figure 5.8. Diet and Order-level microbiome community structure for Donor A.

5.4.2 Dynamic MAP-Model for Prediction of Microbiome Community Structure Changes in Response to Host Diet

There are two steps to generating an MAP-model for predicting a microbial population structure from environmental or microbiome-host interactions, as has been previously presented in detail in Chapter 3. The first step is to generate a community interaction network, using DBN. The second step is to use that network as the architecture for a system of equations that determines the abundance of a bacterial taxon as a function of the environmental parameters and other bacterial taxa that are its parents in the community interaction network.

5.4.2.1 Generate a Community Interaction Network

Normalized and log-transformed community structure data and normalized diet parameters were used to generate an environmental interaction network to identify the possible relationships between taxa and between taxa and host diet in the human microbiome. A CIN was generated as a DBN using BANJO. BANJO, a freely available tool for generating Bayesian Networks (Smith et al. 2006) (https://users.cs.duke.edu/~amink/software/banjo/), was run using the following parameters: ‘Greedy’ searcher, discretization policy of 5 integer values, and a maximum of five parents per node. In addition to input parameters, Diet Parameter nodes were not allowed to have parents in the final network, requiring that diet parameters are always root nodes of the final network and current time point taxa nodes could only have parents from the previous time point’s taxa. The complete set of BANJO parameters are listed in **Appendix B**. The final network is shown in **Figure 5.9**. In the calculated network, all taxa are included but the diet parameters carbohydrates and protein are not found in the final network. In the community interaction network, the diet features of fiber, saturated fats, and sodium are the dietary parameters that have the largest effects, i.e. have the highest connectivity, on microbiome community structure in this individual.

5.4.2.2 Generate MAP Model

The CIN generated in the previous section was used to describe the community interaction network as a system of equations such that the value of every node is a function of the value of its parent nodes. The complete CIN network is found in **Appendix C**. Equations were generated using an Machine Intelligence (MI) evolutionary algorithm for finding the best non-linear equations to fit a dataset ('Eureqa' v 1.2), over 4.2 billion possible solutions were searched to identify the set of mathematic equations that best describe microbiome population structure as a function of current time step's diet and previous time step's population structure. All MAP-model equations are found in **Appendix D**.

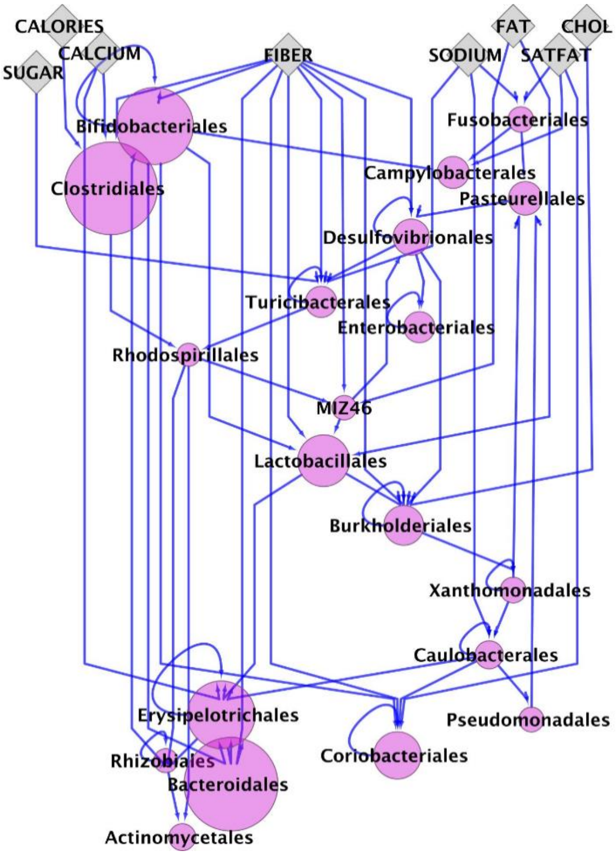


Figure 5.9. MAP-model of human microbiome population structure, predicted as a function of diet parameters. (A) The DBN for community interaction of a gut microbiome identifies how diet parameters and relative abundance of microbial taxa drive changes in the gut microbiome population profile. In this network, top nodes are diet parameters (diamonds), and all other nodes are microbial taxa (circles). Nodes for taxa are sized proportionately to their average abundance across all Donor A microbiomes. Diet parameters (Calories, Sugar, Cholesterol, Fat, Saturated Fat, Fiber, Sodium, and Calcium) were taken from David et al and taxa abundance were considered at the level of Order. Edges between nodes are predicted relationships between taxa and diet or between taxa and other taxa. This network can be used as the basis of an MAP-model, in which the value of any taxa node is described as a function of the values of its parent diet or taxa nodes.

Using the data generated in (David et al. 2014) and our MAP-modeling approach, predicted microbiome community structures significantly correlate with observed microbiome (p-value less than 0.00001, determined by 10,000-iteration bootstrap analysis of results) (**Figure 5.10**). While this prediction appears strong, the microbiome community structures under investigation are relatively stable over the course of these observations. To make sure that the correlation is due to successful application of modeling procedure and not a function of the relative consistency of population structure across observations, computational results were also compared to an 'average-abundance model'. In the average abundance model, the relative abundance of a taxa for any observation is predicted to be equal to the average of observations across all time points. MAP-model prediction is also superior to the average abundance model (p-value 0.00003, determined by 10,000-iteration bootstrap analysis).

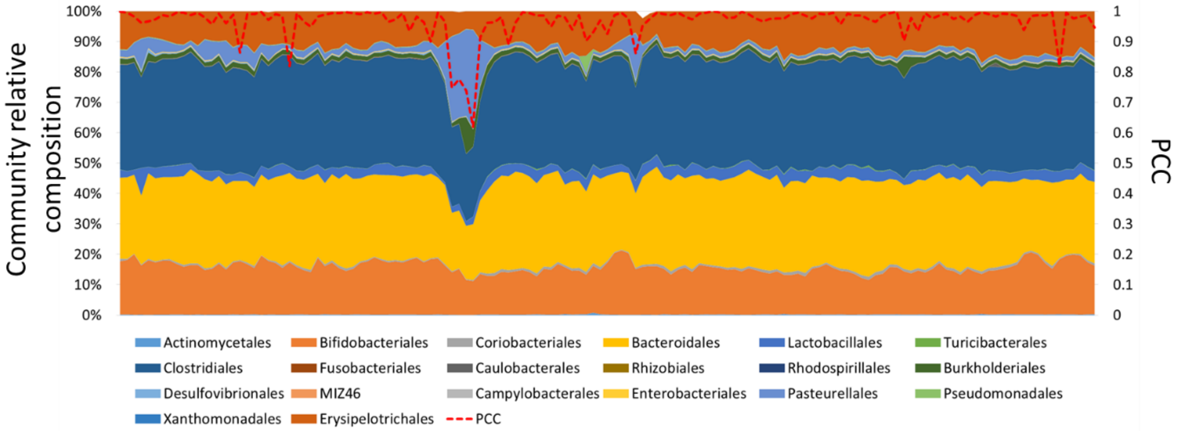


Figure 5.10. Predicted microbiome community structures significant correlates with observed community structures. Note that correlation drops when Donor experiences dysbiosis (about one third of the way through series). This indicated that while dynamic MAP-models can be used to predict microbiome from diet, some environmental perturbations are not included in this model.

5.4.3 Summary of Results

Analysis of the CIN finds that that fiber, sugar, and sodium are the most important drivers of gut microbiome community structure in this dataset for Donor A. Fiber is well known as being crucial to influencing the gut microbiome and promotion of positive gut HMIs (Blaut 2002, Lesmes et al. 2008, Shen et al. 2012). Likewise, dietary sugars have been reported in the prior literature as having a profound influence of microbiome community (Beards et al. 2010, Arora et al. 2012, Shen et al. 2013). The prediction that sodium is a strong driver of microbiome community structure is more intriguing in this context. Sodium is not generally reported as having a strong influence on microbiome. In a 2012 survey of 37 microbiologists with research interests in the effects of diet, disease, and metabolism on microbiomes (http://humanfoodproject.com/guts-germs-and-meals-what-37-microbiologist-say-about-diet/), sodium was ranked low (3.6 of an scale of 1 to 10) in its perceived relevance to microbiomes. Sodium, however, is never present in diet simply as sodium ions, so a question presents itself from these results: What food components are most likely to co-occur with the presence of sodium in the diet? While it is unlikely that sodium itself has a strong influence of microbiome community structure, food preservatives that contain sodium (**Table 5.9**) are known to have a significant effect on the microbiome (McKnight et al. 1999, Chassaing et al. 2015, Lennerz et al. 2015). We propose that it is not sodium in this model that is has a strong influence of microbiome community; rather we hypothesize that it is sodium-containing food preservatives that alter microbiome community composition.

Table 5.9. Food preservatives that contain sodium

Compound Name	Food to Which the Compound Is Added
Sodium acetate	Baked goods, seafood
Sodium benzoate	Beverages, fermented vegetables, jams, fruit fillings, salad dressings
Sodium propionate	Cheese, baked goods
Sodium diacetate	Condiments
Sodium nitrate	Cured meats
Sodium sulfite	Fruit and vegetable products, seafood
Sodium ascorbate	Meat products
Sodium lactate	Meat products
Sodium phosphates	Meat products, cheese, puddings or custards
Sodium erythorbate	Meat, soft drinks
Disodium ethylenediaminetetraacetic acid (EDTA)	Salad dressing, mayonnaise, canned seafood, fruit fillings
Sodium dehydroacetate	Squash

A list of sodium-containing food preservatives, taken from (Doyle and Beuchat 2001).

5.5. Modeling Diet-induced Dysbiosis

The host's diet influences microbiome community structure, and microbiome community can in turn cause dysbiosis in the host. Here, we will combine the models built in Task 1 and Task 2 to predict a microbiome community structure that is a consequence of diet, and then predict whether or not that microbiome community will result in dysbiosis of the host. While there is no opportunity to follow up predictions of diet-induced dysbiosis with experimental data, it is possible to determine if diets that lead to dysbiosis in the model are supported by the available literature. If the model of diet-induced dysbiosis matches well with biological expectation, then we can be confident that the system-scale model of human microbiome is truly extrapolative as well as predictive and that the computational model has captured some key biological elements that underlie HMIs.

Here, we combined MAP-model and SVM for prediction of host dysbiosis from microbiome metabolome models. The complete system-scale modeling approach is summarized **in Figure 5.11**.

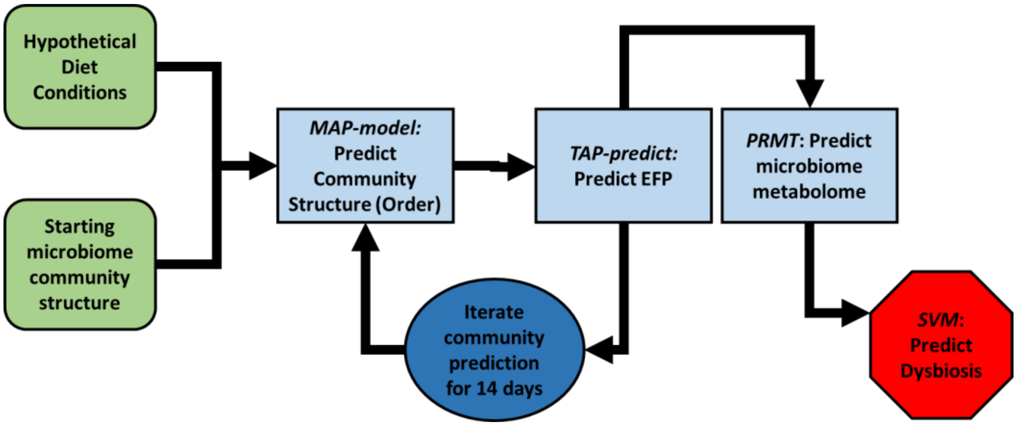

Hypothetical Diet Conditions
Starting microbiome community structure
MAP-model: Predict Community Structure (Order)
TAP-predict: Predict EFP
PRMT: Predict microbiome metabolome
Iterate community prediction for 14 days
SVM: Predict Dysbiosis

Figure 5.11. Summary of system-scale model for diet-induced dysbiosis. The model accepts as input a stating microbiome community structure and a set of diet parameters. Using the dynamic MAP-model approach, the system-scale model predicts the change in microbiome community structure after n days on the proposed diet. The resultant microbiome community is used to generate a prediction of the community enzyme function profile and a community metabolome model. The community metabolome is used as input to an SVM for prediction of host dysbiosis.

5.5.1 Data Selection and Preprocessing

Again, for this Task we return to the David et al. longitudinal microbiome data for Donor A, using diet parameter data and community structure data as the taxonomic level of Order.

5.5.1.1 Non-Independent Diet Parameters: Calories and Carbohydrates

In the set of diet parameters provided by the David et al. study, not all parameters are independent of one another. Calories, for example, is an aggregate measurement derived from a combination of parameters such as dietary sugars, fats, and proteins. Total carbohydrates are likewise a combination of the amounts of sugar and fiber in the diet. While parameters such as calories might ordinarily be estimated as a mathematical function of the known caloric content of the other diet parameters (e.g. dietary sugar, fats, and protein content), the lack of available unit measurements in the provided data make this problematic. The solution chosen here was to use the available dietary data and solve a set of equations such that calories is a function of all other diet parameters and carbohydrates in a function of sugar and fiber. An MI approach was used to determine the best equations to fit observed data, using the same parameters as utilized above.

The identified equations were:

$$\mathit{Calories} = 0.7763 * \mathit{carbohydrate} + 0.0001354 * \mathit{cholesterol} * \mathit{fat} * \mathit{protein} \qquad \textbf{Eq. 5.4}$$

$$\mathit{Carbohydrate} = 0.65 * \mathit{sugar} + 0.271 * \mathit{fiber} \qquad \textbf{Eq. 5.5}$$

Predicted calorie parameters correlates with reported calories with PCC of 0.910. Predicted carbohydrate parameter correlates with reported carbohydrate with a PCC of 0.762.

5.5.2 Predict new Microbiome Communities for Hypothetical Diets

There are three steps for predicting microbiome communities from hypothetical diets: (i) use MAP-model to predict microbiome community structure from diet parameters, (ii) use TAP-predict method to estimate EFP from community structure, and (iii) use PRMT to predict microbiome community metabolomes from EFPs.

5.5.2.1 Predict Microbiome Community Structures: MAP-model

Using the dynamic MAP-model for microbiome community structure prediction as a function, the response of microbiome community to a variety of hypothetical diet conditions was determined. The starting microbiome community structure for all diets was taken from the Day 1 observed microbiome community parameters for Donor A. A total of 80 diets were considered. Diet parameters for calcium, cholesterol, fat, fiber, protein, saturated fat, sodium, and sugar were set to values between 10 and 100 units, in intervals of 10 units, with all other parameters set equal to the average of all diet parameters for Donor A. Diet parameters for carbohydrates and calories were determined as a function of all other diet parameters using equations described in previous section (**Eqs. 5.4** and **5.5**). The MAP-model constructed earlier in this chapter was used to determine microbiome population community structures after modeling 14 days on new diet. Fourteen days was empirically determined to be enough time-steps for microbiome population to achieve a new stable structure in response to diet change. New population structures are found in Appendix and are shown in **Figure 5.12**.

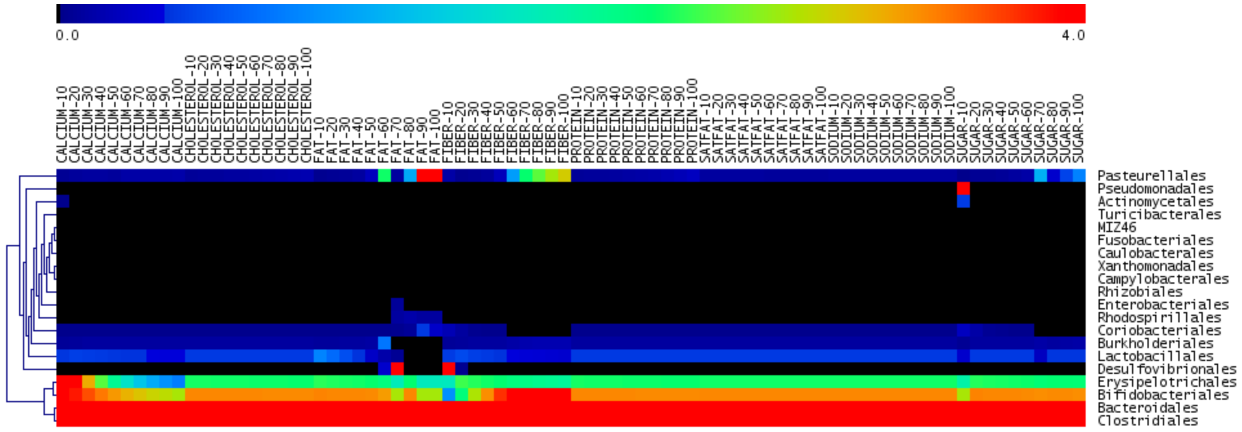


Figure 5.12. Microbiome community structures predicted for 80 hypothetical diet conditions. Population composition is presented as the log2 of relative percent population abundances. Data is hierarchically clustered by population composition, using Euclidian distances.

From **Figure 5.12**, it is observed that *Pateurellales* abundance is increased with increasing fat, fiber, and sugar in diet. *Pseudomondalaes* and *Actinomycetales* are increased in diets very low in sugar. *Erysipelotrichales* and *Bifidobacterales* are increased in diets low in calcium and *Bifidobacterales* are increased in high fiber diets.

In addition to the 80 microbiome community structures generated here, we also incorporated 75 observed Order-level population structures from Donor A: 35 from Dysbiosis and 40 from Healthy (divided evenly between pre- and post-dysbiosis) microbiomes. This subset of observed data will be carried through the subsequent analysis steps described below and will serve as the training data for SVM prediction of dysbiosis from microbiome community metabolome data.

5.5.2.2 Predict Microbiome EFPs: TAP-prediction

Previously in Chapter 5, EFPs were predicted from genus-level taxonomic descriptions of microbiome community structure data. Here, we have used the same approach, applied to the same set of Order-level population descriptions as was used in the dynamic MAP-model of microbiome community.

An Order-level TAP-matrix was generated for the 20 Orders present in the microbiome MAP-model. A total of 1556 genomes belonging to the orders were available for generating the TAP-matrix (**Table 5.7**). No genomes were available for reported bacterium "MIZ46", so for TAP-matrix, the average of the 1331 remaining annotated genomes was used to calculate the average and standard deviations for enzyme function abundances in MIZ46. The resulting EFPs were comprised of 1901 unique enzyme functions.

5.5.2.3 Predict Microbiome Community Metabolomes: PRMT

Using the EFPs generated in the previous section, metabolic models were constructed using PRMT approach for the 80 hypothetical diet microbiomes and for the 75 observed Donor A microbiomes, as initially described in Chapter 3. The predicted CIN had 4370 metabolic interactions between 2864 metabolites mediated by 1305 unique enzyme functions. The metabolites identified as most predictive for dysbiosis earlier in this chapter were all represented in the order-level community metabolic model (**Figure 5.13**).

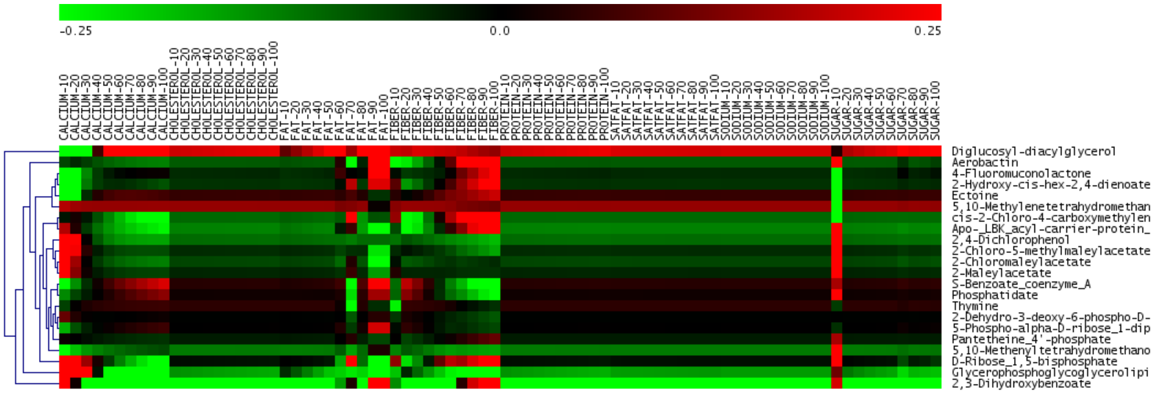


Figure 5.13. Metabolic models (PRMT) for microbiome communities derived from hypothetical host diet conditions. In this figure, only the metabolites previously identified at predictive for dysbiosis are shown. While also calculated, the metabolome of selected 75 observed Dysbiotic and Healthy Donor A microbiomes are not pictured here.

5.5.3 Predict Diet-Induced Dysbiosis Using SVM

With the metabolic model data generated in the previous section, SVM models will be generated to predict which, if any, of the hypothetical diets cause diet-induced dysbiosis in the host. For training models, only those metabolites previously identified as most predictive of host dysbiosis (23 metabolites, **Table 5.6**) will be used.

SVM models were trained using observed Donor A microbiome data (40 Healthy and 25 Dysbiosis microbiomes, as described above) using the same parameters as described in previous sections. SVM trained on observed microbiome data was used to predict which of the hypothetical diet-derived microbiomes are dysbiotic (**Table 5.10**).

Table 5.10. Computational model predicts that 14 days of low carbohydrate or high fat diets induce dysbiosis in host

		10	20	30	40	50	60	70	80	90	100
Diet Parameter	Calcium	N	N	N	**N**	N	N	N	N	N	N
	Cholesterol	N	N	N	**N**	N	N	N	N	N	N
	Fat	N	N	**N**	N	N	N	N	N	D	D
	Fiber	D	N	N	**N**	N	N	N	N	N	N
	Protein	N	N	**N**	N	N	N	N	N	N	N
	Saturated Fat	N	N	**N**	N	N	N	N	N	N	N
	Sodium	N	N	N	**N**	N	N	N	N	N	N
	Sugar	D	N	N	**N**	N	N	N	N	N	N

The changes in microbiome community structure in response to 14 days on different diets were calculated. In novel diet conditions, one diet parameter was set to a value between 10 and 100 while all other diet parameters were set to a value equal to the average diet parameter across all observations. For each resulting microbiome community structure, EFP and Metabolome were determined. Host dysbiosis was predicted from microbiome metabolome. In these predictions, a diet low in fiber, low sugar, and high fat induced dysbiosis in host. Diet conditions resulting in dysbiosis are 'D's highlighted with **red** text and light red backgrounds and non-dysbiosis microbiomes are 'N'. Diet conditions that approximate observed average diet conditions are highlighted in **blue** text.

5.5.4 Summary of Results

Two different diet conditions are linked to dysbiosis in this predictive model: diets high in fat and diets low in the carbohydrates sugar and fiber. This matches well with prior observations and a diet high in fat is reliably associated with dysbiosis (Schulz et al. 2014, Murphy et al. 2015, Camilleri 2016, Zhang and Yang 2016). Low fiber intake is also associated with dysbiosis in this prediction. This too matches well with expectations. Diets very low in sugar were predicted to induce host dysbiosis, which seems less intuitive, but is a prediction that actually also fits quite well with prior published observations. While a diet low in sugar (specifically, low relative to the typical 'Western' diet) undoubtedly has health benefits to a human host, starving the gut microbiome for fermentable sugars decreases microbiome diversity

(Beards et al. 2010, Arora et al. 2012, Shen et al. 2012). In the model generated here, we propose that the predicted link between very low sugar diets and dysbiosis is due to this reduction in diversity.

One of the opportunities that derive from this analysis is to compare TAP-predictions for microbiome communities described at the level of order and ant the level of genus. One relevant question is how similar are the EFPs for the same populations where predicted from different taxonomic levels. For the 75 microbiome community microbiomes selected for analysis, the average PCC between order and genus-level EFP predictions is 0.872. This is a good correlation, but these results indicate that EFPs generated for the same microbiome communities are not identical when they are predicted from different taxonomic-level descriptions of population structure. If there was a known metagenome for these communities, then the question could be asked, which EFP is a closer to the biological observation. Unfortunately, for this dataset there is no metagenomic data available. However, this question will be explicitly addressed in Chapter 6, using gnotobiotic mouse microbiome data.

6. GENERATING A SYSTEM-SCALE MODEL OF A MOUSE OBESITY HOST-MICROBIOME INTERACTIONS

While the importance of the microbiome to human health is undeniable, the opportunity to analyze the human microbiome *in situ* is fraught with difficulty. Humans are genetically diverse, influencing their predisposition to diseases, and the way they interact with their microbiome. They also encounter a wide variety of environments and other humans, making their microbiomes a continually evolving, dynamic community. This mutability makes it difficult to peel back the layers of interactions and identify the specific role the microbiome plays in a given human's phenotype. A powerful laboratory system that enables hypothesis-driven experiments on the interactions between host, microbiome, and environment is the gnotobiotic mouse. Gnotobiotic mice, from 'gnostos' meaning known and 'bios' meaning life, are laboratory animals with a completely characterized microbiomes and therefore are ideal for laboratory experiments linking microbiome to host phenotype.

Previous chapters have used available data to predict HMI interactions, but all prior efforts in this study have come up against the same limitations: the availability of data and the opportunity to validate model predictions in a controlled laboratory environment. Both the Pseudomonas-host interaction and human microbiome studies have been observational in nature and did not include experimentally manipulated parameters.

In this Chapter, the analysis methods developed in previous chapters will be leveraged for the analysis of a set of gnotobiotic laboratory animal studies of HMI mechanisms. New analyses will explicitly address the weaknesses identified in previous chapters. Multiple datatypes and data sources will be utilized to generate an integrated system of computational models to link host diet, microbiome functional capacity, and host phenotype for an investigation into the microbiome's role in host obesity.

Results will be applied to predicting specific diet approaches for reducing host propensity to obesity as a function of microbiome community structure.

6.1 Background

From a 2014 study, over a quarter of the world's population is overweight and over 8% are obese (Sonnenburg and Backhed 2016). Obesity results in a significantly increased risk in mortality (**Figure 6.1**) (Global et al. 2016), and obesity is associated with elevated risk for serious health conditions such as hypertension, type 2 diabetes, heart disease, stroke, osteoarthritis, and some cancers (Consortia 1998, Kasen et al. 2008, Bhaskaran et al. 2014, Ryan and Heaner 2014). In estimates from 2008, obesity-related medical costs were $147 billion in the U.S. (Finkelstein et al. 2009), with an additional $3.38 to $6.38 billion in lost productivity. It has been predicted that by 2030 obesity-related medical costs in the U.S. could rise by an *additional* $48 to $66 billion per year (Wang et al. 2011). Even a small reduction in obesity in the U.S. could lead to billions in reduced medical costs and prevent tens of thousands of premature deaths.

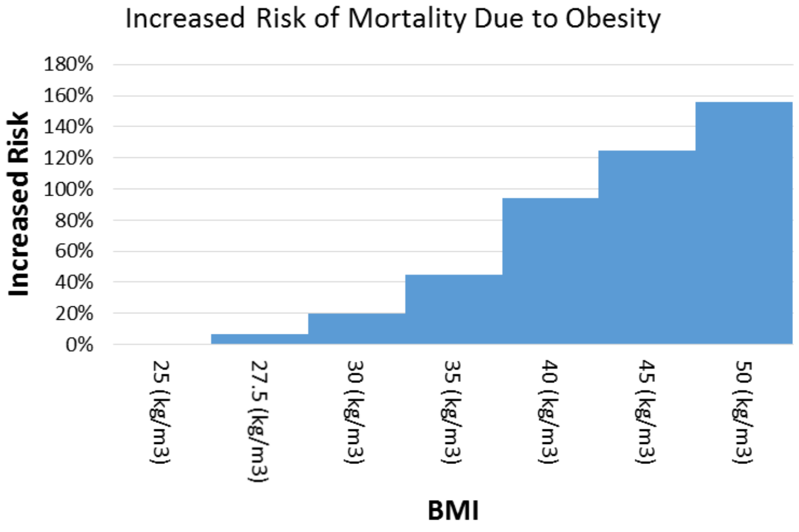


Figure 6.1. Risk of mortality increases with BMI. Increased risk of mortality is calculated relative to a BMI of 25 kg/m^3, considered to be a healthy weight

6.1.1 The Microbiome and Obesity

The microbiome is known to play a significant role in obesity. Understanding the relationship between obesity and the microbiome will be a powerful tool for combating obesity in the U.S. and around the world. In a meta-analysis of microbiome obesity association studies collected by (Walters et al. 2014), the bacterial taxa in microbiomes associated with obesity were compiled from multiple studies (**Table 6.1**). The most salient feature of this table is that no result in which a bacterial taxa was identified as obesity-related has been duplicated in an alternate study and in one case, the same bacteria has been associated with both lean and obese microbiomes in different studies. One previously reported common feature to most studies is that the relative abundances of *Bacteroidetes* are decreased and *Fimicutes* are increased in the microbiome of obese individuals (Turnbaugh et al. 2009a, Walters et al. 2014). However, even this commonly held view has been questioned by other experimental observations (Duncan et al. 2008, Schwiertz et al. 2010).

The summary of microbiome analyses indicates that obesity markers in the microbiome have highly multivariate features that may be dependent upon a wide range of factors not explicitly controlled in these studies of human subjects. It may not be achievable to assemble a sufficient body of human microbiome data available for analysis due to the tremendous inherent variability in the human populations and the great many factors that contribute to obesity. In order to address the questions of how the microbiome predisposes a host to obesity, it is necessary to turn to experimental animal models.

Table 6.1. **Bacterial taxa that have been linked to obesity by previously published analyses**

Taxa	Inc	Dec	References
Actinobacteria			
Bifidobacterium (genus)		🟥	(Schwiertz et al. 2010)
Bifidobacterium animalis		🟥	(Brandt and Aroniadis 2013)
Euryarchaeota			
Methanobrevibacter smithii	🟥	🟥	(Schwiertz et al. 2010, Patil et al. 2012)
Firmicutes			
Oscillospira [sp]		🟥	(Tims et al. 2013)
Clostridium cluster XIVa	🟥		(Verdam et al. 2013)
Roseburia intestinalis	🟥		(Tims et al. 2013, Verdam et al. 2013)
Eubacterium rectale	🟥		(Furet et al. 2010, Ferrer et al. 2013, Tims et al. 2013)
Faecalibacterium prausnitzii		🟥	(Furet et al. 2010, Verdam et al. 2013)
Lactobacillus (genus)	🟥		(Collado et al. 2007, Bervoets et al. 2013)
Lactobacillus casei/paracasei		🟥	(Brandt and Aroniadis 2013)
Lactobacillus reuteri	🟥		(Brandt and Aroniadis 2013)
Bacteroidetes			
Bacteroides (genus)	🟥		(Patil et al. 2012, Tims et al. 2013)
Bacteroides vulgates		🟥	(Bervoets et al. 2013, Verdam et al. 2013)
Bacteroides uniforms		🟥	(Verdam et al. 2013)
Alistipes (genus)		🟥	(Verdam et al. 2013)

Cells are highlighted in red if bacteria listed in left column has been associated with an Increase ("Inc") or Decrease ("Dec") in abundance in obesity according the to references in the right column. Note that is no case has a link between a change in bacterial taxa abundance and host obesity been replicated consistently across studies.

6.1.2 Gnotobiotic Animal Models for Studying HMIs

A significant difficulty in the analysis of human HMIs is the great diversity of the human microbiome. Another is the ethical barriers preventing experimental manipulation of human subjects that would enable hypothesis-driven laboratory experiments in HMI. To overcome these difficulties, we turn to gnotobiotic laboratory animal models. Gnotobiotics is the use of animal models where the microbiome community is either completely known or else tightly controlled (Ward and Trexler 1958).

A variety of host-microbiome systems have been used to identify correlations between abundances of specific microbiome taxa and host phenotype and to molecular mechanisms of HMIs (Kostic et al. 2013). The Hawaiian bobtailed squid, for example, is a well-studied HMI in which the host squid actively acquire *V. fisheri*, a luminescent bacterium, from their environment and cultivate that bacteria in special light organs within the squid (Nyholm et al. 2000, Nyholm and McFall-Ngai 2003, 2004). Drosophila melanogaster, the fruit fly and reliable laboratory model animal, has been used to study the role of their comparatively simple microbiomes on microbial parthenogenesis and effect on host innate immunity (Dionne and Schneider 2008, O'Callaghan and Vergunst 2010). While zebrafish have been used as a model of vertebrate HMIs (McFall-Ngai 2007, Maynard et al. 2012), a more common tool for laboratory models of HMI is the mouse (Spor et al. 2011, Walters et al. 2014).

The use of the mouse model has many significant advantages (Bouskill et al. 2011). The set of available genetic tools for mouse genome manipulation and the range of available genotypes and phenotypes of laboratory mouse make this animal model highly appealing for hypothesis-driven experiments. The mouse model also comes with a deep background in murine immunology, genetics, and gastroenterology owing to its long and common use in scientific and medical research. Previously published experiments have shown that microbiome communities collected from human donors can be successfully transplanted into germ-free mice and that the microbiome community is capable of establishing itself in its new host (Turnbaugh et al. 2006, Yi and Li 2012, Kostic et al. 2013).

One previously published gnotobiotic mouse experiment, by (Ridaura et al. 2013), provides a specific and highly relevant example of the utility of using gnotobiotic mice as a model system for human HMIs (**Figure 6.2**). In this experiment, microbiomes were collected from twins where one twin was obese and the other non-obese. Collected microbiomes were transplanted into germ-free mice and, once it was established that the transplanted communities had taken up residence in their new hosts, it was determined that mice that had received a transplant from an obese human donor were themselves prone to obesity **(Figure 6.2**). This experiment demonstrates clearly the advantages of using the mouse model, specifically (i) a human microbiome can be transplanted into a mouse gut, and (ii) the resulting mouse phenotype (i.e. propensity for obesity) is representative of the human phenotype from which the microbiome was collected.

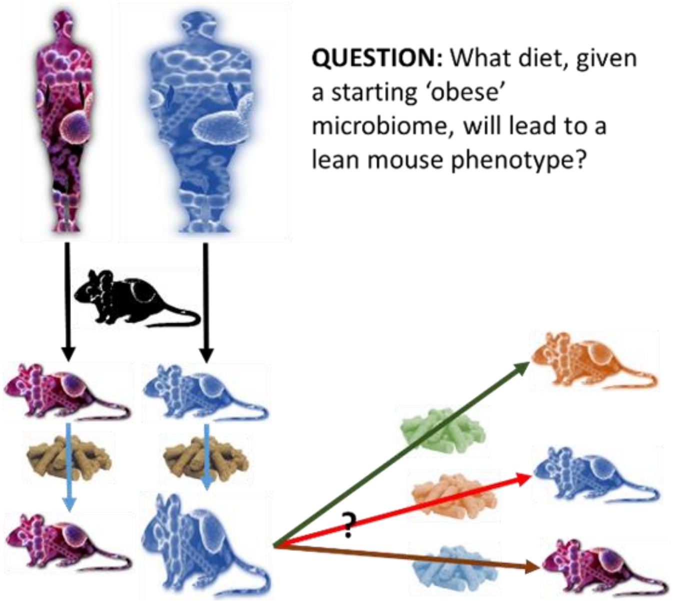

QUESTION: What diet, given a starting 'obese' microbiome, will lead to a lean mouse phenotype?
?

Figure 6.2. Microbiome community can dramatically change how otherwise similar hosts will respond to the same environmental conditions. In a previously published experiment, microbiomes from lean (red) and obese (blue) twins were transplanted into germ-free mice (black mouse). In this experiment, it was found that the 'Obese' microbiome leads to an Obese phenotype, 'Lean' microbiome leads to lean phenotype in mice given the same diet (tan pellets). Our goal is to use predictive and metabolomic modeling to identify a diet (colored pellets) that will restore a Lean phenotype to mice with 'Obese' microbiomes.

6.1.3. Knowledge Gaps and Innovation

A recent meta-analysis of microbiome studies for prediction of host obesity concluded that although subjects could be classified as lean or obese by their microbiomes within a study with significant accuracy, signatures of obesity were not consistent between studies (Walters et al. 2014). A key knowledge gap that this Chapter seeks to overcome is to develop a computational model that successfully predicts obesity from microbiome community data across multiple datasets and numerous experimenters.

This will be achieved through several technical innovations. First, significant improvements will be made to the TAP-prediction tool for estimated EFPs from microbiome community structures. This will be accomplished through a statistical evaluation of thousands of published sequenced and annotated genomes and a stochastic approach to modifying the TAP-matrix within biologically determined dynamic ranges to optimize predictions of EFP using a large database of mouse microbiome metagenomes. Second, the dynamic MAP-model approach will be further refined to make accurate predictions of dynamic microbiome communities both as a function of the initial microbiome community structure and diet parameters, but also interactions between bacteria in the final microbiome community structure. Third, we will apply a novel Machine Intelligence approach for prediction of host obesity from a microbiome community that is robust across multiple experimental treatments, starting microbiome community structures, and laboratories.

6.2. Outline of Experimental Approaches

The goal of this chapter is to combine tools and biological understanding accumulated in Chapters 4 and 5 with a novel dynamic microbiome population model developed in this aim. The collected microbiome and phenotype data investigating the relationships between microbiome, diet, and obesity in a gnotobiotic mouse model will serve as the training and validation sets from this analysis. The combined models will enable the following predictions:

1. Predict new microbiome population structure given an initial microbiome population and diet.
2. From a predicted microbiome, predict the microbiome metabolome.
3. From a predicted metabolome, predict if that microbiome will lead to an obese or lean mouse host phenotype.

This combined set of models will be used in a GA approach that will predict the best diet for selecting lean mouse host phenotype for any arbitrary starting metagenome population structure. Additional significant advancements to the procedure, specifically the inclusion of community metabolome to the dynamic model, are also proposed.

This Chapter is divided into four closely linked Tasks:

Task 1: Generate a dynamic model of the mouse microbiome community structure. This Task builds upon the MAP-model approach to generate predictive models of the final mouse microbiome community structure that results from a consequence of the host's diet and initial microbiome community structure.

Task 2: Predict microbiome community enzyme function profiles from community structures. This Task makes significant improvements to the TAP-prediction methods for estimating metagenomes from microbiome community structure by taking advantage of an available dataset of known mouse microbiome community structures and metagenomes.

Task 3: Predict Obesity from microbiome metabolomes. This Task enables prediction of mouse propensity for obesity resulting from interactions with the gut microbiome. The measure of success in this Task will be creating a predictive model of microbiome-induces obesity that is robust across multiple experimental datasets.

Task 4: Construct a system-scale computational model of mouse gut HMI. In this Task, the individual computational models build and validated in Tasks 1 through 3 will be combined into a single, system-scale model of mouse obesity HMI. This model is called the **i**ntegrated **M**icrobiome models of **O**besity for the **U**nderstanding of **S**ystem-scale **E**mergent properties (**iMOUSE**). The iMOUSE model will enable in silico experiments for mouse obesity HMI and will be used to propose diets that maximize a lean HMI phenotype conditional upon the starting microbiome community structure.

6.3 Selected Microbiome Datasets

Each task of this chapter draws from a different mouse microbiome published study. Four different microbiome datasets were utilized in the aim, drawing from a wide variety of experimental sources to construct the different computational models proposed in this aim. Datasets were integrated such that all experimental data could be described using a uniform set taxonomic descriptors for community structure and refer to a common set of host dietary parameters. For simplicity, the datasets are referred to as ‘Transplant’, ‘Gradient’ ‘Obesity’ and ‘Catalog’.

6.3.1 “Gut Microbiome from Twins Discordant for Obesity Modulate Metabolism in Mice”

The manuscript by Ridaura et al. (Ridaura et al. 2013) describes an experiment in which the obesity phenotype of human donors is transferred into mice via a microbiome transplant. In this study, germ-free C57BL/6J male mice were inoculated with microbiome communities collected from twins discordant for obesity. Mice that received an ‘Obese Microbiome’ gained more weight, even on a LF diet, than mice that received a ‘Lean Microbiome’ transplant. Microbiome community structures for Lean- and Obese-microbiomes were collected from the Supplemental Files of this manuscript. This dataset is used primarily as an opportunity to validate the iMOUSE model.

This dataset is referred to as the ‘Transplant’ data in subsequent analyses.

6.3.2 “A Catalog of the Mouse Gut Metagenome”

A mouse gut microbiome dataset “A catalog of the mouse gut metagenome” (Xiao et al. 2015), was used for this Task. This data set is comprised of 184 mouse gut microbiomes with paired metagenome and population abundance data. Microbiomes were collected from a wide variety of mouse strains (8 strains) that were maintained at seven different housing labs/facilities (**Table 6.2**). 68% of the mice in the dataset are male and 74% were raised on a LF diet. Data was available through the GigaDatabase website, companion website for the *GigaScience* journal (http://gigadb.org/dataset/100114). There were 67 bacterial Classes in the community structure data and1558 unique enzyme functions (EC annotations) present in the available metagenomic data. Results from this study indicated that mouse provider and housing conditions had a marked effect on microbiome community structure and functional representation in the microbiome metagenome. This dataset is used to improve our ability to predict EFPs from microbiome community compositions.

This dataset is referred to as the ‘**Catalog’** data in subsequent analyses.

Table 6.2. Distribution of lab location, mouse strain, mouse gender, and diet in ‘Catalog’ dataset

Housing Lab		Strain		Gender		Diet	
DTU	24	**CB7BL/6**	**100**	Male	126	High Fat	44
NIFES	35	CV129	34	Female	58	Low Fat	140
UCPH	10	BALB/c	8				
Pfizer	80	SJL	8				
BGI	15	129S	8				
CMR	20	SJL-CB75BL/6	8				
		NOD	8				
		Swiss Webster	10				

Housing Labs abbreviations are DTU: The Danish Technical University, NIFES: National Institute of Nutrition and Seafood Research, UCPH: University of Copenhagen, BGI: University of Chinese Academy of Sciences, and CMR: CMR Facility. Note that strain C57BL/6, highlighted in **bold**, is the most common strain in this study and is also the strain of lab mouse used in the other available datasets (i.e. Gradient, Obesity, and Transplant).

6.3.3 "Diet Dominates Host Genotype in Shaping the Murine Gut Microbiota"

A dramatic shift in mouse microbiome community structures have been observed between mice on HF and LF diets. In the manuscript by Carmody et al. (Carmody et al. 2015), authors hypothesize that the large changes in community structure are due to the significant differences between the HF and LF diet conditions. In an experiment presented in this manuscript, mice were given diets representing a gradient between HF and LF diet conditions. Adult male C57BL/6J mice raised on LF diets were fed mixed LF and HF-diet pellets in proportions of 0, 1, 10, 25, 50, 75, and 100% HF diet for seven days. Data were collected from 33 mice from the initial microbiome communities and again after seven days on the new diet for a total of 66 microbiome community observations. Data was collected from MG-RAST (http://metagenomics.anl.gov/) using the 'MGRASTer' tool (https://github.com/braithwaite/MGRASTer/) in R. This data was used here to generate predictive models for the effects of diet of microbiome community structures.

This dataset is referred to as the '**Gradient'** dataset in subsequent experiments.

6.3.4 "Data and Analysis of Diet-induced and Obesity-Associated Alterations of Gut Microbiota of 129S/Sv and C57BL/6J Mice"

Changes in microbiome community structure that is simply in response to host's diet and changes in microbiome community structure that promote obesity in the host are difficult to distinguish in mouse microbiomes. In a dataset published by Xiao et al. (Xiao et al. 2017) (http://gigadb.org/dataset/100271), this difficulty is addressed by considering two different mouse genotypes. In mouse strain C57BL/6J (BL6), treatment of mice with a cyclooxygenase (COX) inhibitor prevents HF-diet induced obesity. In mouse strain129S6/SvEvTac (Sv129), treatment with a COX inhibitor accentuates HF-diet induced obesity. Data was available through the GigaDatabase website, companion website for the *GigaScience* journal (http://gigadb.org/dataset/100114). In the study by Xiao, the host's diet was found to be the principle driver of the microbiome community and no strong relationship between obesity in the host and microbiome was uncovered. This data is comprised of 54 microbiome community structures and were used primarily to generate computational models that predict host obesity from emergent properties of microbiome community.

This dataset is referred to as the ‘**Obesity’** dataset in subsequent experiments.

6.3.5 Describe Microbiome Experimental Using a Common Set of Identifiers

In order to integrate the selected microbiome manuscripts into a single, cohesive dataset suitable for meta-analysis, all experimental results must first be described using the same set of microbiome community and experimental condition identifiers.

6.3.5.1 Select Taxonomic levels for Microbiome Community Structure Descriptions

In order to integrate the selected datasets, all microbiome communities need to be described using a common set of bacterial taxonomic identifiers. Twenty taxa (4 Orders, 15 Genera, and one category for ‘Other’) were selected. The class ‘Other’ comprised on average less than 3% of the total microbiome populations for ‘Gradient’ data set. In the ‘Catalog’ dataset, the category ‘Other’ averages less than 13% of population composition on average and less than 10% in the ‘Transplant’ dataset.

The complete set of taxa used in this Task is listed in **Table 6.3**.

Table 6.3. Bacterial taxa used to describe mouse microbiome community structures

Taxa	**Taxonomic Level**	**Brief Description**
Anaerostipes	Genus	*Anaerostipes* is anaerobic, Gram-positive, and occurs in the human gut.
Bacteroides	Genus	*Bacteroides* species commonly found in the human gut, where they play a fundamental role in processing of complex carbohydrates.
Blautia	Genus	*Blautia* are common in the human gut microbiome and produce acetate. IBS patients have increased levels of *Blautia* species.
Butyrivibrio	Genus	*Butyrivibrio* are common in the gastrointestinal systems of many plant-eating animals.
Clostridium	Genus	*Clostridium* are Gram-positive bacteria, and includes the diarrhea-causing *Clostridium difficile*.
Collinsella	Genus	The abundance of *Collinsella* correlate strongly with high levels of inflammatory compounds.
Erysipelotrichaceae	Genus	Erysipelotrichaceae increase abundance with a high-fat diet and are associated with inflammation-related disorders of the gastrointestinal tract.
Eubacterium	Genus	*Eubacterium* are common in the gut microbiome and help to digest resistant starches.
Lachnospiraceae	Genus	The *Lachnospiraceae* are an anaerobic bacteria found in the human gut. Members of this family are linked to obesity and may protect against colon cancer in humans by producing butyric acid.
Lactobacillus	Genus	*Lactobacillus* are Gram-positive, facultative anaerobes or microaerophilic and are commonly found in the gut microbiome.
Oribacterium	Genus	*Oribacterium* are found in higher abundance in the gut microbiome with high-fat diets and are potentially linked to inflammation.
Parabacteroides	Genus	*Parabacteroides* help digest high-fiber diets and their levels are elevated in the presence of resistant starches.
Porphyromonas	Genus	*Porphyromonas* are Gram-negative obligate anaerobes. Some species are associated with autoimmune diseases.

Ruminococcaceae	Genus	*Ruminococcaceae* are common bacteria in the gut microbiome and help to digest resistant starches. *Ruminococcaceae* increase in abundance with a diet high in plant starches.
Ruminococcus	Genus	*Ruminococcus* are Gram-positive gut anaerobes commonly found in gut microbiome. They help digest resistance starches and are associated with reduced risk of diabetes and colon cancer.
Clostridiales	Order	*Clostridia* are obligate anaerobes. They are commonly found in animal microbiomes and some can be pathogens.
Atopobium	Order	*Atopobium* are Gram-positive anaerobes.
Desulfotomaculum	Order	*Desulfotomaculum* are sulfate-reducing, obligate anaerobes. *Desulfotomaculum* can cause food spoilage in poorly processed canned foods.
Lactococcus	Order	*Lactococcus* produce lactic acid as the sole product of glucose fermentation.
OTHER	N/A	

Bacterial description information was collected from the NCBI Taxonomy Browser (https://www.ncbi.nlm.nih.gov/taxonomy).

6.3.5.2 Define Mouse Host Diet Parameters

Similar to how diet was described as a vector of nutrient parameters for modeling effects of diet of microbiome community in Chapter 5, it is necessary to be able to describe LF and HF diets as a vector of nutrient parameters for modeling mouse microbiome community dynamics. Low Fat (LF) diet parameters were collected from available data sheets for ENVIGO "Teklad Custom Diet" (http://www.envigo.com/products-services/teklad/laboratory-animal-diets/), comprised of Diet Mix TD.08811 made with Mineral Mix TD.94046 and Vitamin Mix TD.94047. High Fat (HF) data parameters were collected from available datasheets for LabDiet "JL Rat and Mouse/Auto 6F (http://www.labdiet.com/). Additionally, the amino acid composition for casein in LF diet was inferred from an analysis found in Gordon et al (Gordon et al. 1949). The complete set of HF and LF diet parameters are listed in **Table 6.4**. For use in this analysis, all diet parameters were normalized to arbitrary values between 20 and 80 and $\log_2$ transformed.

Table 6.4. Mouse diet compositions for High Fat (HF) and Low Fat (LF)

	Nutrient (g/Kg)	**HF**	**LF**
	Protein	19.3	18
Amino Acids	Ile	0.89	0.87
	Leu	1.73	1.52
	Lys	1.51	0.97
	Met+Cys	0.62	0.98
	Phe+Tyr	1.97	1.41
	Thr	0.78	0.68
	Val	1.09	0.9
	Trp	0.20	0.23
	His	0.49	0.44
	Ala	0.53	1.13
	Arg	0.65	1.03
	Asp	1.47	1.87
	Glu	4.25	4.52
	Gly	0.33	0.94
	Pro	1.81	1.53
	Ser	1.08	0.98
Carbohydrates	Carbohydrate	50.34	39.79
	Starch	11.7	38.9
	Glucose	0	0.12
	Fructose	0	0.15
	Sucrose	34.84	0.62
	Lactose	3.8	0
	Fiber (cellulose)	5	15
Fats	Total Fat	23.2	6.2
	Saturated	14.15	1.24
	Mono saturated	7.192	1.37
	Poly unsaturated	1.856	0.24

	Nutrient (g/Kg)	**HF**	**LF**
Minerals	Calcium	1.5351	1.17
	Potassium	1.347534	0.66
	Magnesium	0.10449	0.22
	Iron	0.026058	0.038
	Zinc	0.007095	0.0085
	Magnesium	0.002709	0.016
	Copper	0.001333	0.0011
	Iodine	0.000043	0.00021
Vitamins	Niacin	0.0057	0.009
	Panthothenate	0.00304	0.0037
	Pyridoxine	0.00133	0.001
	Riboflavin	0.00114	0.0009
	Folic acid	0.00038	0.00019
	Biotin	0.000038	0.00003
	Vit B12	0.00475	0.005
	Vit E	0.0285	0.0045
	Vit A	0.00152	0.002
	Vit D3	0.00038	0.00043
	Vit K	0.000143	0.002
	Kcal/g	**4.7**	**3.17**

6.4 Predict Changes in Microbiome Community Structures in Response to Host Diet: Dynamic MAP-Model

The goal of this Task is to create a computational model that predicts the final mouse microbiome community structure that results as a function of the microbiome initial population structure and host's diet. This will be accomplished by making significant improvements to the dynamic MAP-model approach, such that diet, initial microbiome community structure, and predicted community bacterial interactions are used in prediction. The model will provide two principle tools to provide insights into HMI. The first is a CIN for microbiome community and host diet, which will highlight specific relationships between taxa and diet. The second is a predictive MAP-model that can be integrated into a system-scale model of mouse obesity HMI.

As before, generation of a MAP-model requires two consecutive steps. The first step is to generate a Community Interaction Network (CIN) using Bayesian Network Inference. Here, the CIN will predict microbiome community interactions as nutrient parameters, initial microbiome community structure, and final microbiome community structure. The second step in the MAP-model procedure is to transform the CIN into a system of integrated mathematical equations, such that the value of every child node is a function of the values of its parent nodes. For this Task, the CIN-based equations will be solved as linear functions using a least squares estimate. This will enable an additional form of information to be extracted from the CIN, specifically whether the relationship between a parent and a child node is positive or negative (i.e. does an increase in the value/abundance of a parent node lead to an increase or a decrease in the abundance of a child node).

6.4.1 Selection of 'Gradient' Mouse Microbiome Dataset

This Task utilizes the 'Gradient' dataset. In this dataset, each microbiome observation is comprised of the following data: diet parameters, initial microbiome community structure and final microbiome community structure. Diet parameters are comprised of 50 diet features (**Table 6.3**). Initial and final microbiome community structures are comprised of 20 taxonomic bacterial abundances (**Table 6.4**).

6.4.2 Generate CIN from 'Gradient' data

As described in previous sections, CIN for interactions between mouse microbiome and diet were generated as DBN. DBN were generated using BANJO (Smith et al. 2006) (https://users.cs.duke.edu/~amink/software/banjo/documentation/) using the same parameters as in previous sections. In addition, the following restrictions were made. No node is permitted to the parent of a diet parameter node or a prior time step's taxa and prior time steps taxa can only be parents of current time step's taxa. The complete set of BANJO parameters can be found in **Appendix E**.

A visualization of the mouse microbiome CIN can be found in **Figure 6.3** and the CIN network itself in **Appendix F**. In the interaction network, 46% of nutrient parameters are amino acids, 13% are carbohydrates, 4% are fats, 13% are minerals, and 25% are vitamins. Relative to the distribution of nutrient types in the total set of diet parameters, network diet nodes are significantly enriched for vitamin nutrient features (calculated as hypergeometric mean, p-value less than 0.05). The bacteria that have the greatest influence of population structure (i.e. the most child nodes in network) are *Parabacteroides* and *Butyrivibrio*. The bacterial nodes most regulated by other community interactions are *Desulfotomaculum*, *Ruminococcus*, *Clostridium*, and 'Other'. Only *Bacteroides* and *Lactobacillus* have no direct parent nodes that are nutrient parameters. *Porphyromonas*, the most abundant bacteria in the mouse microbiome in this dataset, has only nutrient parameter parents and no predicted interaction with other taxa. *Ruminococcus*, which is known to be associated with digestion of complex carbohydrates in the microbiome (Ze et al. 2012), is positively affected by the nutrient parameter 'Fiber' in the interaction network. The only bacteria taxa associated with fat intake in the network is *Eubacterium*, and

Eubacterium have been previously shown to have an advantage in the microbiome in high fat, high sugar diets (Turnbaugh et al. 2008).

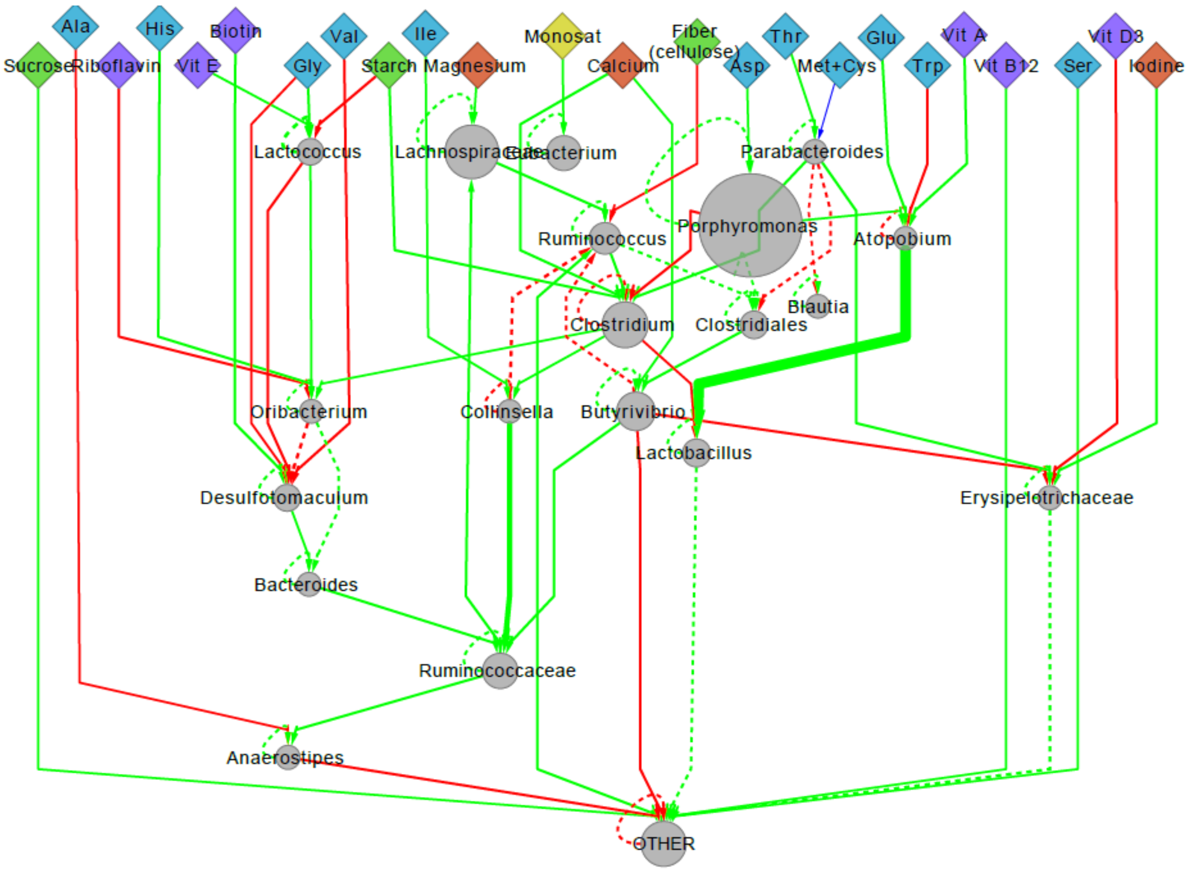


Figure 6.3. Mouse diet-gut microbiome interaction network. This figure combines the host-microbiome CIN with results from the dynamic MAP-model. Diamonds are diet parameters, colored by the nutrient category (Table X). Circles are bacterial taxa and are sized by their average relative abundance across all Donor A microbiomes. Edge thickness is proportionate to the absolute value of edge weight in the map-Model. Edges are green when edge weight is positive and red when negative.

6.4.3 Generate Dynamic MAP-Models from ‘Gradient’ Microbiome Data

The second step of the dynamic MAP-model approach is to use the CIN network to determine a system of equations that can be used to generate a prediction model of a microbiome community structure as a function of environmental factors. For this model, MAP-model, we will consider the relative abundance of a taxa in microbiome community to be:

$$\mathbf{MAP-model}: \quad taxa_i^t = \sum_{j=1}^{Diet_{Parents\,of\,i}} w_{j,i} diet_j^t + \sum_{k=1}^{Taxa_{Parents\,of\,i}} w_{k,i} taxa_t^t + \sum_{l=1}^{Taxa(t-1)_{Parents\,of\,i}} w_{l,i} taxa_l^{t-1} + c_i \qquad \textbf{Eq. 6.1}$$

Where $taxa_i^t$ is the relative abundance of taxa i at time t, $Diet_{Parents\,of\,i}$ is the set of diet parameters that are parents of taxa *i* in CIN, $Taxa_{Parents\,of\,i}$ is the set of taxa abundances that are parents of taxa i in CIN, $Taxa(t-1)_{Parents\,of\,i}$ is the set of taxa abundances from previous time point, t-1, that are parents of taxa *i* in CIN, $w_{x,i}$ is a weight between CIN node *x* and taxa i, and c_i is a constant value associated with taxa *i*

In addition to the dynamic MAP-model, two additional methods for prediction of microbiome community structure were considered. In these approaches, microbiome community structures were calculated as functions of *all* diet and/or *all* initial microbiome taxa without consideration of the relationships inferred by the CIN.   The goal of these methods is to determine how much, if any, predictive value is conferred to the model by incorporation of the CIN for microbiome population structure. We hypothesize that models built using the CIN will demonstrate a higher accuracy in predicting microbiome community structure than models that do not use the CIN.

Diet Only: $$taxa_i^t = \sum_{j=1}^{AllDiet} w_{j,i} diet_j^t + e_i$$ **Eq. 6.2**

Diet and Data: $$taxa_i^t = \sum_{j=1}^{AllDiet} w_{j,i} diet_j^t + \sum_{l=1}^{AllTaxa^{t-1}} w_{l,i} taxa_l^{t-1} + e_i$$ **Eq. 6.3**

Where *AllDiet* and *AllTaxa*$^{t\text{-}1}$ indicate that **all** nutrient parameters and taxa relative abundances, not just those that are parent nodes in the CIN, are considered as parameters in these equations.

Data were randomly divided into training (23 microbiome pairs) and validation (9 microbiome pairs) subsets. The three models of microbiome community ('Map-model', 'Diet Only', and 'Diet and Taxa') were solved using least squares estimate (QR decomposition of matrix in R). The same training and validation subsets were used for each model. Predictive power of models was determined by the Pearson Correlation Coefficient between the predicted and observed population structures (as $\log_2$ relative abundances) for both training and validation subsets. The resulting edge weights from equations inform edge weights in **Figure 6.3**. Significance of predictions were calculated using a bootstrap approach (10,000 iterations). Results for training and validation subset are summarized in **Figure 6.4**.

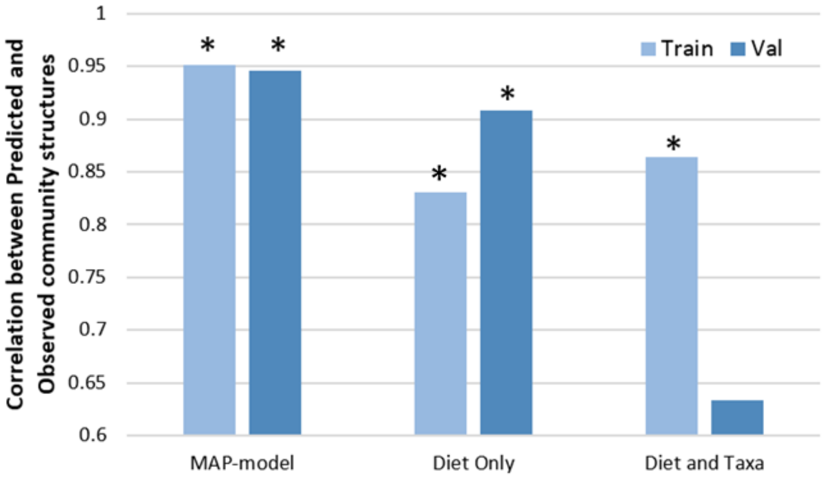


Figure 6.4. Results for mouse microbiome community predictions. Three modeling approaches were considered: “MAP-model’, “Diet Only”, and “Diet and Taxa”. Predictions that performed statistically significantly better than random (p-value less than 0.05) are indicated with an “*”.

6.4.4 Summary of Results

‘Diet and Taxa’ model has good results for the training set, but the poor results for validation suggest that this approach is vulnerable to overfitting, possibly due to the relatively small sample size. ‘Diet’ only model has a good result for the validation set. It may be hypothesized that this is because the model is constructed from an experiment where population structures at the initial time point are fairly similar, therefore including taxa at the initial time point does not add that much new and relevant information to the model. ‘MAP-model’ has the best predictive results for both the training and validation datasets with significant PCC of about 0.95 for validation data subset. This strong prediction accuracy provides evidence that using this model in subsequent analysis steps will likely result in biologically meaningful results.

By combining the most predictive MAP-model results with the CIN, additional insights into the community interactions can be gained. More than 70% of the interaction types are positive in nature and the average weight of positive edges is greater than the average weight of negative edges, although difference is not statistically significant.

6.5 Optimize Prediction of EFP from Community Structure Data: TAvP-prediction

Previously, we have utilized a method for predicting microbiome's Enzyme Function Profile from community structure (Chapter 5). When applied to the 'Catalog' database of microbiomes and metagenomes (as described in detail below), this approach generates a set of predicted metagenomes that correlates with the observed metagenomes with a median Spearman's Correlation of 0.88. This compares favorably with the Tax4Fun method which reports a median Spearman's correlation of 0.75 for their best approach for mammalian gut metagenome predictions (Asshauer et al. 2015). While by this metric TAvP-prediction considerably outperforms prior published results, a direct comparison between methods is complicated by the use different datasets, different starting datatypes, and different ontologies for metagenome annotations. At the very least, it can be confidently stated that the TAP-prediction method is of comparable accuracy to other published metagenome prediction methods.

Results from Chapter 5 provide additional insight into opportunities to improve the TAP-prediction methods. In that chapter, EFPs were predicted from microbiome community structures described at the taxonomic level of Order and at the level of Genus. One relevant question is how similar are the EFPs for the same populations when predicted from different taxonomic levels. For the 75 microbiome community microbiomes selected for analysis, the average PCC between order and genus-level EFP predictions is 0.872. This is a good correlation, but EFP-prediction from different taxonomic levels results indicate that EFPs for the same microbiome communities are not identical when they are predicted from different taxonomic-level descriptions of population structure.

There remains considerable opportunity for improvement in the TAP-prediction of mouse gut EFPs from community structures. In this task, an approach to incorporate a deeper statistical analysis of published sequenced genomes and observed mouse gut microbiome metagenomes into the prediction of EFPs from community structure data is proposed. These efforts resulted in a substantial increase in EFP prediction accuracy over previous TAP-prediction results.

6.5.1 Selection of ‘Catalog’ Mouse Microbiome Data

This Task utilized the dataset from ‘Catalog’ mouse microbiome experiment (Xiao et al. 2015), which is comprised of 184 paired microbiome shotgun metagenome and population structure data collected from a range of different mouse strains housed in different research labs and with different diet regimes. Having a set of known metagenomes, by which the accuracy of metagenomic predictions from community structure can be tested, provides an ideal opportunity to make significant improvements to the initial version of the metagenome TAP-prediction tool introduced in Chapter 3 and used in Chapter 5.

For this analysis, data were randomly divided into sets of 122 training microbiomes and 62 validation microbiomes.

6.5.2 Describe Multiple Approaches for Predicting EFPs from Microbiome Community Structures

The initial method, first discussed in Chapter 3, predicts enzyme function abundance in TAP-prediction is calculated as the following:

$$EC_i^n = \sum_{j=1}^{\boldsymbol{Taxa}} AveEC_i^j * Taxa_j^n \qquad \text{Eq. 6.4}$$

Where EC_i^n is the abundance of enzyme function *i* in microbiome *n*, ***Taxa*** is the set of bacterial taxa reported present in the microbiome, $AveEC_i^j$ is the average number of genes for enzyme function *i* in taxa j, and $Taxa_j^n$ is the relative abundance of taxa *j* in microbiome *n*.

While this method has served well in previous analyses, there is an opportunity for improvement. For example, the numbers of genes for an enzyme function attributed to a taxa in this method is derived from the average of gene abundances for genomes in a published database. It is unlikely that the distribution of bacteria within a given taxa in the database is identical to the distribution found in the microbiome community. Also, using only the average gene abundance for a function in a taxa provides no information about the possible distribution of gene abundance counts across the genomes that represent a given bacterial taxa. For example, two separate bacterial taxa both might have an average of five copies for ‘glucose-fructose oxidoreductase’ genes per genome. But if one taxon has a standard deviation in abundance of 0.2 and the other a standard deviation of 7.0, then an equivalent abundance of these two taxa in different microbiome populations will have very different consequences on the expected variation in glucose-fructose oxidoreductase in the resulting EFP.

The novel modifications described here, named Taxonomic Average and Variance Profile prediction (TAvP-prediction), will not only utilize the average abundances of gene functions within a taxonomic grouping, but also the distribution in order to increase the accuracy of TAP-predictions by fitting prediction results to observed metagenomes.

Three methods for fitting predicted EFP to observed EFP were considered: Average Difference, TAvP-Uniform, and TAvP-Boltzmann. Each approach is described in detail below. Additionally, the unmodified TAP-prediction approach will be used as a control.

6.5.2.1 Average Difference

The simplest approach to leveraging observed metagenomes to optimize TAP EFP predictions is to add a simple error term to the initial TAP-prediction method:

$$EC_i^n = \left(\sum_{j=1}^{\boldsymbol{Taxa}} AveEC_i^j * Taxa_j^n \right) + \boldsymbol{e_i} \qquad \textbf{Eq. 6.5}$$

Where e_i is calculated as the difference between the average of predicted enzyme function abundance across a set of metagenomes and the average of observed enzyme function abundance. In this fashion, a single error term is added to each enzyme function abundance prediction in order to improve the correlation between predicted and observed EFPs.

6.5.2.2 Taxonomic Average and Variance Profile Prediction (TAvP-prediction)

The alternative approach, TAvP-prediction, considers not only the average abundance of an enzyme function in a taxonomically grouped set of sequences bacterial genomes, but also the spread of observed enzyme functions counts from the mean. An alternative approach is to add an error term to each average enzyme function for each taxa in the microbiome:

$$EC_i^n = \sum_{j=1}^{\boldsymbol{Taxa}} \left(AveEC_i^j + \boldsymbol{e_j^i} \right) * Taxa_j^n \qquad \textbf{Eq. 6.6}$$

Where e_j^i is an error term that is added to the average enzyme function abundance for enzyme activity *i* in bacterial taxa *j*. This error term is calculated such that the range of possible values is determined from the observed distribution in published genomes for the number of genes for enzyme function *i* in taxa *j*. These taxa and enzyme-specific terms can be identified for an entire TAP-matrix using a stochastic approach.

The TAvP-prediction approach for optimizing the TAP-matrix is described using the following pseudocode:

```
Pseudocode for TAvP-matrix optimization:

Given:      Pop = Matrix of microbiome population structures
            EFPobs = EFP for Pop microbiomes, collected from
            observed/published metagenomics data
            EFPrange = the dynamic ranges between min and max values
            for each enzyme function
            TFCave, TFCsd = Initial Taxonomic Function Count averages
            and standard deviations.

Function PREDICT_EFC(Pop, TFC): returns a predicted EFP given a
   microbiome population structure and a TAP-matrix

Function TWEAK(taxa, enzyme, TFCave, TFCsd): Modifies enzyme function
   count for taxa by an amount informed by TFCsd.

Function PCC(EFC1, EFC2) Returns PCC between EFCs

Function SelectEnzyme(EFPrange): If linear approach, return enzyme
   with uniform probability.  If Boltzman approach, return enzyme,
   preferentially selecting those that have larger dynamic ranges
   across observed metagenomes

Function SelectTaxa(enzyme, TFCsd): If linear approach, return taxa
   with uniform probability.  If Boltzman approach, return taxa,
   preferentially selecting those that have large SD for enzyme

bestEFP = PREDICT_EFC(Pop, TFCave)
bestPCC = PCC(EFPobs, EFCinitial)
bestTFC = TFCave

For i = 1 to NumInterations

     Enzyme = SelectEnzyme(EFPrange)
     Taxa = SelectTaxa(Enzyme, TCFsd)
```

```
      newTFC = TWEAK(taxa, enzyme, TFCave, TFCsd)
      newEFP = PREDICT_EFP(Pop, newTFC)
      newPCC = PCC(newEFP, bestEFP)
      if newPCC > bestPCC then
           bestPCC = newPCC
           bestTFC = newTFC
           bestEFP = newEFP

Output bestEFP, bestPCC

END
```

There are two versions of the function '**TWEAK**' for this approach: Uniform and Boltzmann. In the Uniform method, each taxa and each enzyme is randomly selected with equal probability. In the Boltzmann method, selection of enzyme function is weighted such that enzyme function abundances that are highly variable across microbiomes, and such that taxa that have higher standard deviations for the abundance of that function are selected more frequently. This weighting of selections was done using the Boltzmann distribution (McQuarrie 2000):

$$p_i = \frac{e^{\varepsilon_i/kT}}{\sum_{j=1}^{M} e^{\varepsilon_j/kT}} \qquad \textbf{Eq. 6.7}$$

Where p_i is the probability of selecting enzyme function or taxa i. ε_i is the equal to the standard deviation of function abundances. kT is the Boltzmann constant multiplied by the temperature of the system. M is the total possible number of either enzyme functions or taxa in the system.

6.5.3 Compare EFP Prediction Methods

The methods TAvP-Uniform and TAvP-Boltzmann were each run using one million iterations. The number of times each taxon and each enzyme were 'Tweaked' (**Figures 6.5 and 6.6**) and the running 'BestPCC' (**Figure 6.7**) were collected during the million iterations for both methods.

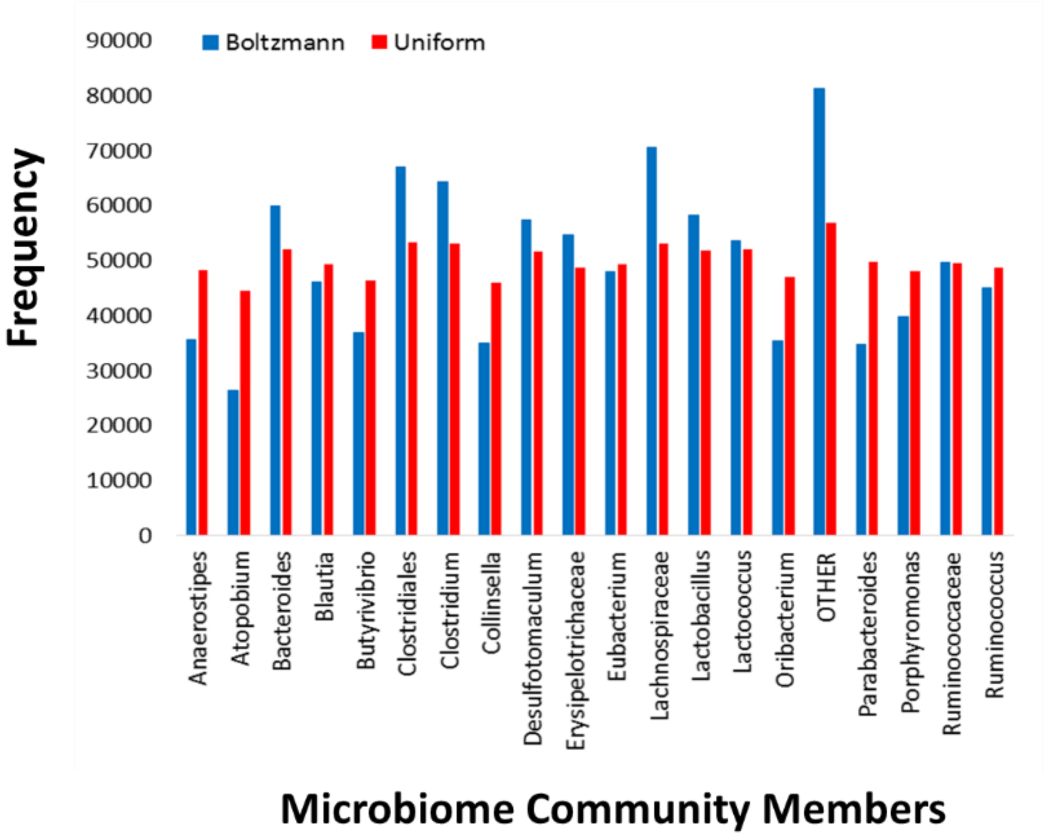


Figure 6.5. Frequency at which specific enzymes and taxa were selected across 1 million iterations. Graphs show that 'Boltzmann' approach lead to very different frequencies for selection for enzymes and taxon relative to the 'Uniform' approach. The most frequently selected taxa by the Boltzmann approach was the general bin for 'Other'

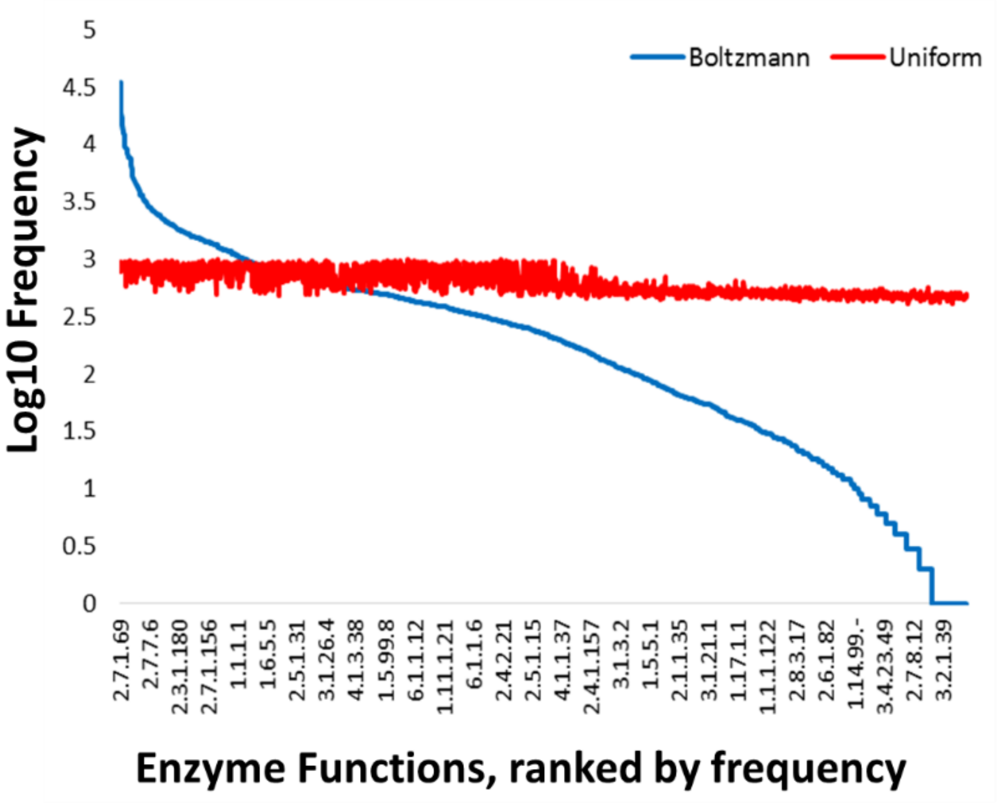


Figure 6.6. Frequency at which specific enzymes were selected across 1 million iterations. Graphs show that the 'Boltzmann' approach lead to very different frequency of selection for enzymes and taxa relative to the 'Uniform' approach. Enzymes are ranked in this figure on the x-axis from most frequently modified to least. The most selected enzyme function by the Boltzmann approach was 2.7.1.69, a sugar transporting phosphotransferase system. The least selected was 5.99.1.4, a 2-hydroxychromene-2-carboxylate isomerase.

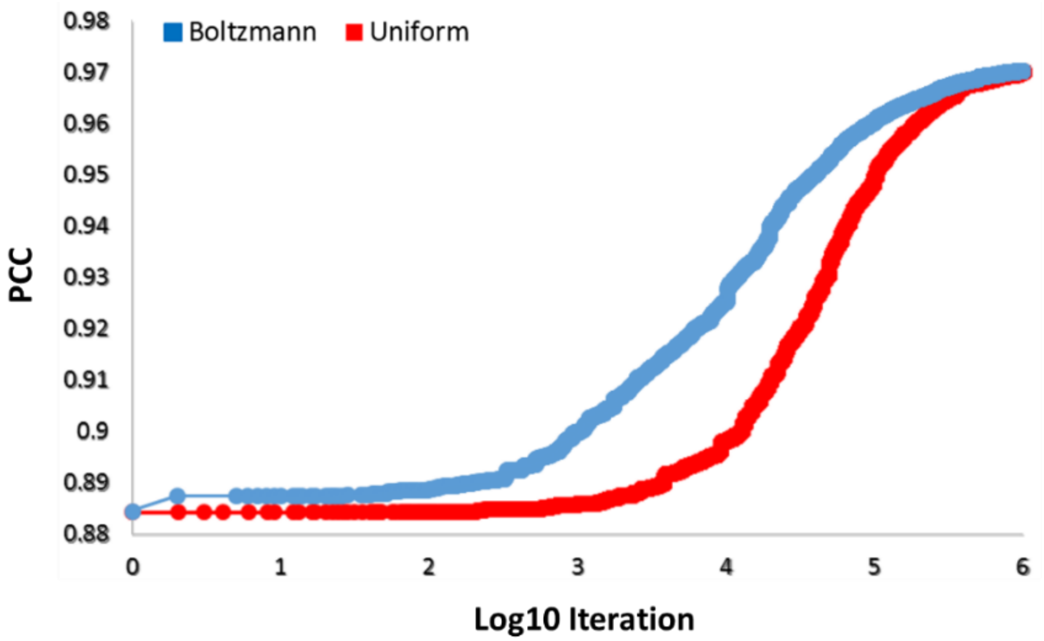


Figure 6.7. Tracking increasing best correlations over iterations for 'Uniform' and 'Boltzmann' approaches. In this figure, log10 iteration is on x-axis and correlation between observed and predicted EFP is on y-axis.

6.5.3.1 Differences between Uniform and Boltzmann approaches

From the results, the difference between Uniform and Boltzmann selection approaches is apparent. The Boltzmann approach resulted in a difference of about three-fold difference between the frequency of the most and least frequently selected taxa (**Figure 6.5**). The differences between frequencies of selection for enzyme activity, however, are far more dramatic. There is a difference of over four orders of magnitude between the most and least frequently selected enzyme function (**Figure 6.6**). The Uniform method, as expected, selects taxa and enzyme function with fairly uniform frequency. While the TAP-matrices optimized using the Boltzmann approach show a considerable lead in EFP prediction for the great majority of iterations, the Uniform selection catches up with the Boltzmann method towards the end of one million iterations (**Figure 6.7**). At the end of TAP-matrix optimization, both the Boltzmann and the Uniform methods yield TAP-matrices that predict EFPs with strong correlations with the observed EFPs (PCC 0.9702 for Boltzmann, 0.9703 for Uniform).

From these results, an interesting question arises: While Boltzmann and Uniform optimization methods lead to nearly identical EFP predictions, do they do so by arriving at the same TAP-matrix? The results will suggest very different possible interpretations of this approach. If the two methods arrive at different TAP-matrices, then perhaps the method arrives at an effective mathematical solution, but not one that necessarily reflects the distribution of metabolic and enzymatic functions across species in the microbiome. If both methods arrive at similar TAP-matrices, then the probability that the best computational solution is also reflective of a true biological state is increased. To determine the similarity between MAP-matrices, pair-wise correlations between the initial Boltzmann-optimized and Uniform-optimized were calculated (**Table 6.5**).

Table 6.5. Correlations between EFP predictions based on initial and optimized TAP-matrices

	Initial	**TAvP-Uniform**	**TAvP-Boltzman**
Initial	1		
TAvP-Uniform	0.905	1	
TAvP-Boltzmann	0.903	0.984	1

Results in **Table 6.5** show that Boltzmann and Uniform are very similar to one another (PCC 0.984) and are approximated equally different from the initial TAP-matrix. This demonstrates that optimization methods arrive at very similar TAP-matrix solutions, enhancing the expectations that these matrices represent an insight into the underlying biological distribution of metabolic functions in the mouse gut microbiome, in addition to being a computationally effective approach for improving metagenome predictions from microbiome community data.

6.5.3.2 Application to Validation Subset Microbiomes

So far, only the training set data has been considered for the accuracy of EFP predictions from microbiome community structures. To determine the utility of the optimized TAP-matrices and TAvP-prediction for metagenomes outside the training set, and to compare these results with the initial and Average Difference control approach, the 62 community structures and metagenome data set aside for validation will be considered. The results of predictions, as correlations between predicted and observed EFPs, are shown in **Figure 6.8**.

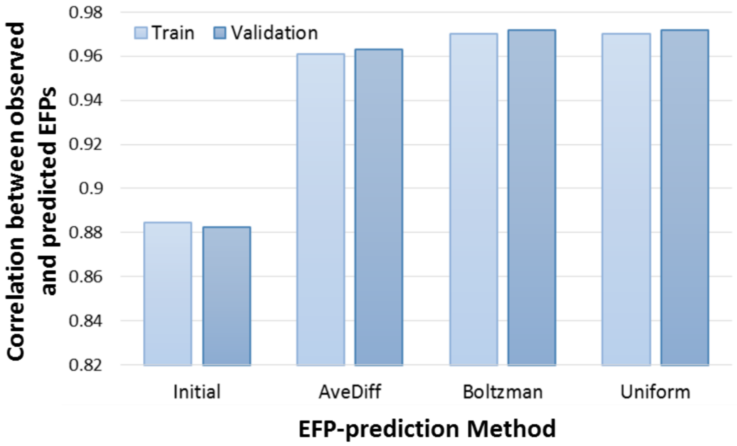


Figure 6.8. Correlations between observed EFP and predicted EFPs generated using different metagenome prediction methods.

Results for EFPs predictions show that all methods, Average Difference, TAvP-Boltzmann, and TAvP-Uniform, show considerable increases in the accuracy of EFP predictions from community structures over the initial TAP-prediction approach. Average Difference approach actually performs quite well, resulting in a predicted EFP that correlated with the observed EFP with a PCC of 0.963 for the validation subset. However, TAvP-predictions, using either Boltzmann or Uniform optimized TAP-matrices perform even better, correlation with biological observations with a PCC of 0.972. Boltzmann or Uniform approach results were nearly identical in outcomes, with both training and validation PCC results differing only at the ten thousandths decimal place.

6.5.4 Summary of Results

Here, we have demonstrated a method, TAvP-predictions, that results in a substantial (13%) increase in the accuracy of predicting metagenomes from community structure. Accuracy of predictions is not only much improved over our own previously published method, TAP-prediction, but accuracy is also easily equivalent or better that the top most-effective metagenome-predicting tools present in the available literature (Larsen et al. 2015b, Shoaie et al. 2015). The stochastic approach for fitting EFP predictions using observed EFPs and a statistical analysis of thousands of previously sequenced and annotated genomes represents a significant innovation over alternate available prediction tools.

While both Boltzmann and Uniform-optimized TAP-matrices yielded nearly identical results in their accuracy of metagenome prediction from community structure, only one can be chosen for use in the subsequent analysis steps here. In the remaining analyses in this chapter, TAvP-predictions using Boltzmann-optimized TAP-matrix will be used.

6.6 Predict Host Obesity as Function of Microbiome Community

When it comes to the relationship between diet, microbiome community, and obesity, it can be difficult to separate out cause from effects. Observed changes in the microbiome community structure are due only to a change in host diet and which changes in microbiome structure are influencing host obesity cannot be easily disentangled. To address these questions, we turn to the published 'Obesity' dataset (Xiao et al. 2017). In this study, genetic variants of mice that are conditionally either resistant or susceptible to obesity in response to a HF diet are considered in order to deconvolute the diet and obesity-specific changes in microbiome community.

In the conclusions reported by Xiao et al. (Xiao et al. 2017), they state that "[t]he changes in the composition of the gut microbiota were predominantly driven by high-fat feeding rather than reflecting the obese state of the mice". They also report that in their analysis that the abundance of butyrate and propionate producing bacteria in microbiome may "at least in part contribute to" the differences between obese and non-obese mice.

In previous analysis (Chapter 5 as well as (Larsen et al. 2015a, Larsen and Dai 2015)), we have reported that microbiome community metabolome is more predictive of host dysbiosis than microbiome community structure. We anticipate that our approach will positively associate host's obesity with emergent properties of the microbiome where the Xiao et al analysis could not. In this Task, we will generate predictive models of both host obesity and HF-diet from microbiome community data.

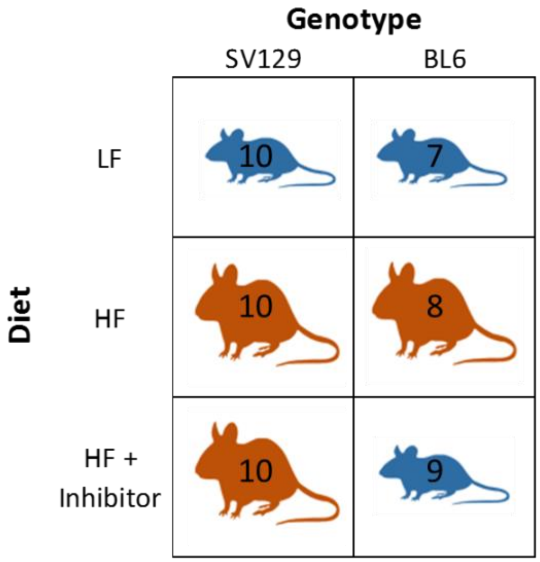


Figure 6.9. Relationships between genotype, diet, and obesity in two laboratory mouse strains. Two mouse genotypes, SV129 and BL6, have different responses to a high fat (HF) diet in the presence of a COX-inhibitor. In the figure, little **blue** mice indicate a Lean phenotype and large **orange** mice indicate an Obese phenotype. The numbers on the mouse indicates the number of microbiomes for that experimental condition in the available data.

6.5.1 Selected of 'Obesity' and 'Transplant' Microbiome Datasets

The 'Obesity' dataset, which uses different mouse genotypes to distinguish the relationship between diet and microbiome and the relationship between microbiome and host obesity, is utilized in this analysis (**Figure 6.9**). The 54 'Obesity' dataset microbiomes were divided into subsets for model training (36 microbiomes) and validation (18 microbiomes). In addition, a further validation condition was included: the two microbiomes derived from human microbiome transplants in the 'Transplant' dataset. From the published experimental observations in the 'Transplant' dataset, an 'Obese-microbiome' transplanted community is expected to confer an Obese phenotype of host and the 'Lean-microbiome' transplanted microbiome a Lean phenotype. Microbiome community structures are represented in **Figure 6.10** and analyzed by hierarchical clustering in **Figure 6.11**. Effective predictions of host obesity have to work not only on microbiomes from the 'Obesity' dataset, but also have to be effective in extrapolating to microbiome community data gathered from other available sources such as the 'Transplant' dataset.

6.6.2 Predict Emergent Properties of Mouse Microbiome 'Obesity' and 'Gradient' Communities

Using the TAvP-prediction approach and the Boltzmann TAP-matrix calculated in Task 2, and described in the previous section, and the PRMT-score metabolic modeling approach described in Chapter 3, 'Obese' and 'Transplant' data community structure data were used to predict EFPs for all microbiomes (**Figure 6.10**). EFPs were used to calculate PRMT-score metabolic models (**Figure 6.11**).

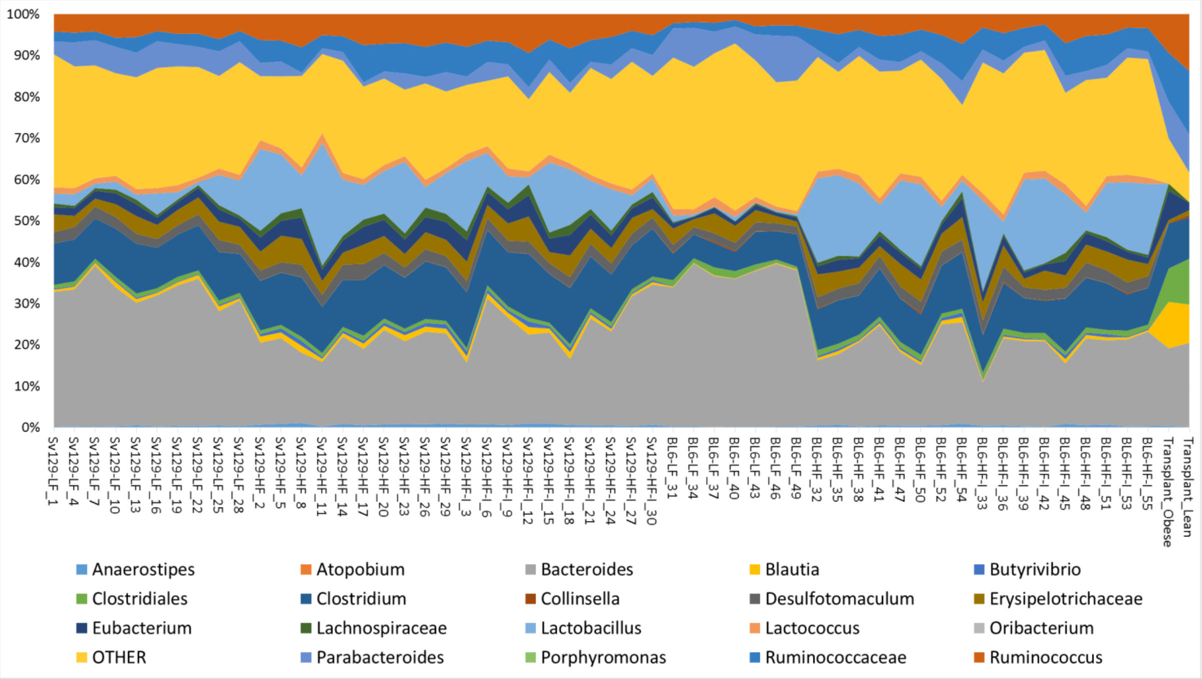


Figure 6.10. 'Obesity' and 'Transplant' microbiome community structures.

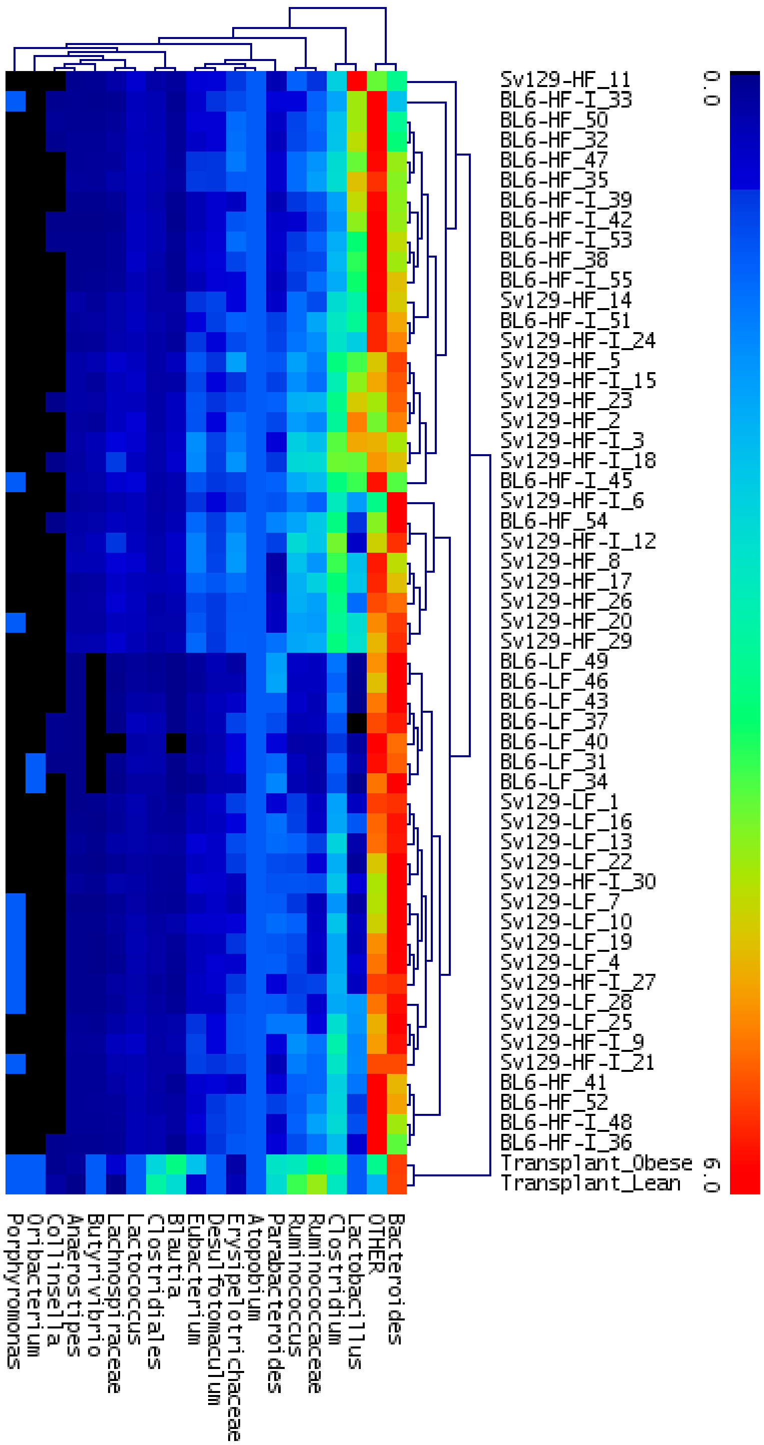


Figure 6.11. Hierarchical cluster microbiome community structures for 'Obesity' and 'Transplant' datasets

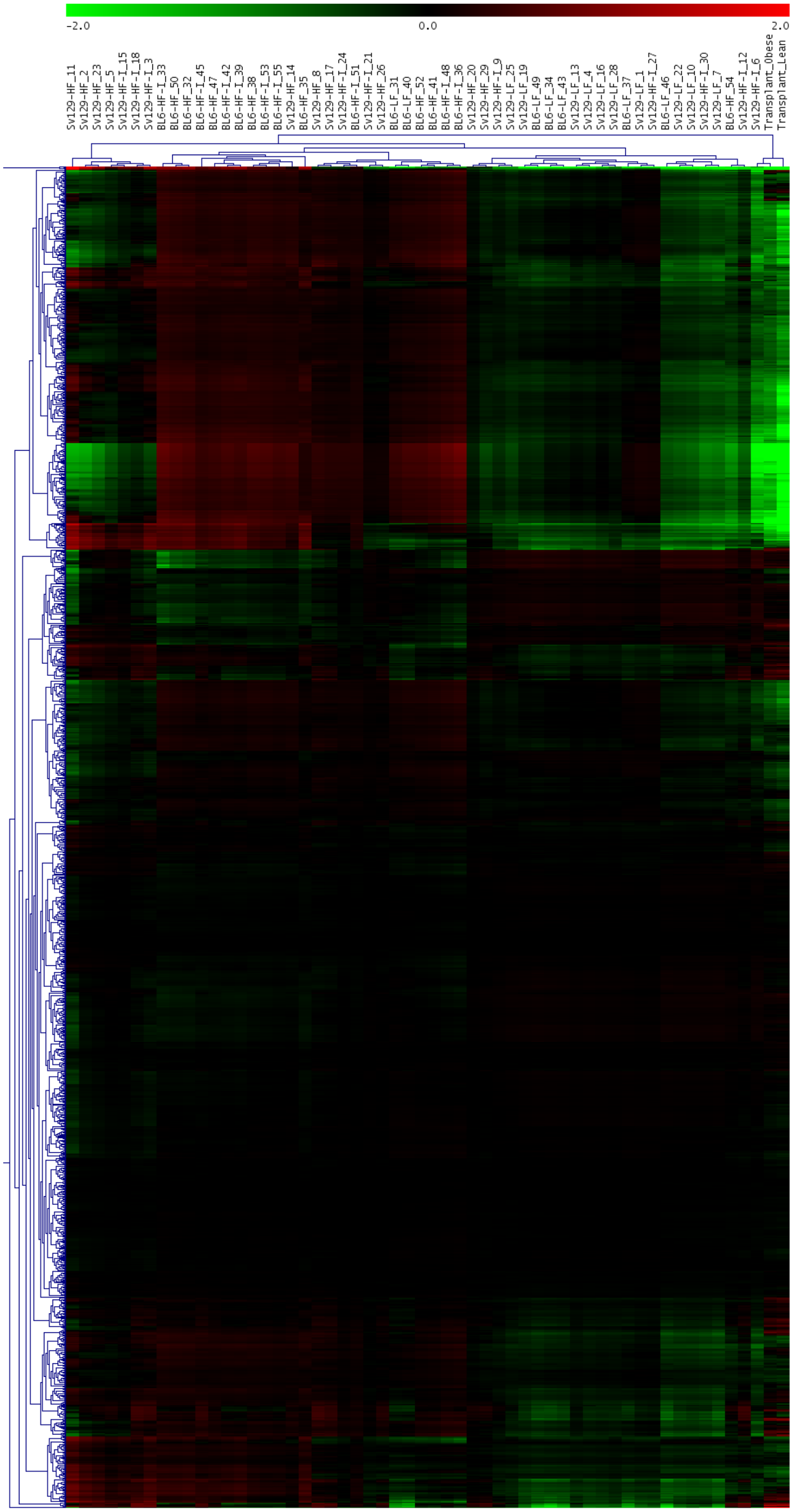


Figure 6.12. Hierarchical cluster EFP for 'Obesity' and 'Transplant' datasets.

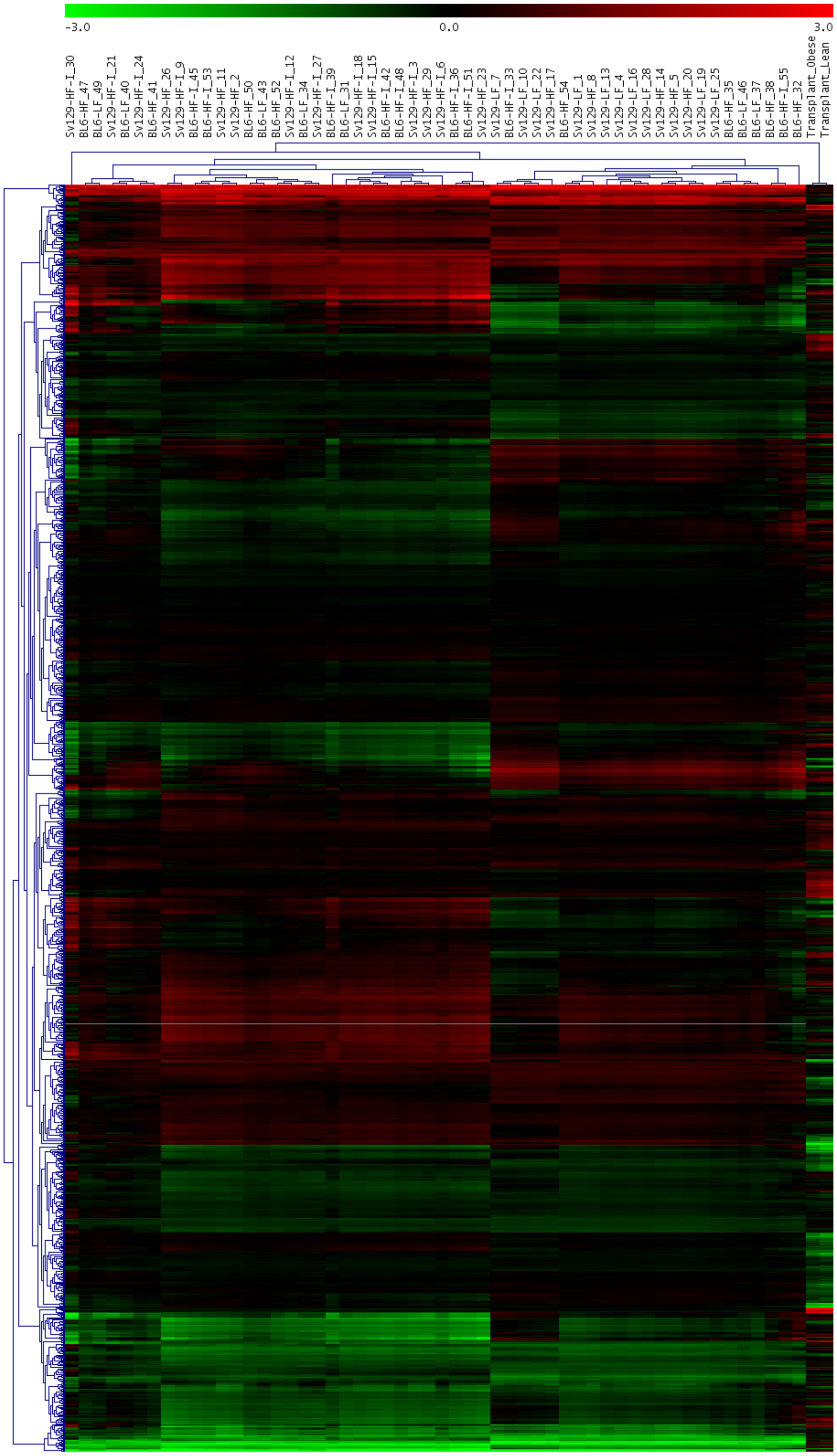


Figure 6.13. Hierarchical cluster PRMT-scores for ‘Obesity’ and ‘Transplant’ datasets.

From the clustering images of microbiome data, patterns emerge. In the cluster of ‘Obesity’ and ‘Transplant’ microbiome community population structure data (**Figure 6.9**), microbiomes cluster by diet, HF or LF, with ‘Transplant’ microbiomes clustering separately. Within diet types, microbiomes cluster by mouse strain. When microbiomes are clustered by EFP (**Figure 6.10**), more clusters become apparent with Sv129 HF, BL6 HF with and without COX inhibitor treatment, and LF diets regardless of mouse strain form more distinct clusters. When clustered by EFP, ‘Transplant’ microbiomes remain outliers. When microbiomes are clustered by PRMT-scores (**Figure 6.11**), microbiomes are most likely to cluster by mouse strain, but not by phenotype (obese or lean). Clustering by PRMT highlights that ‘Transplant’ microbiomes are very distinct from ‘Obesity’ microbiomes more strongly than when microbiomes are clusters by either community structure or EFP. Under no conditions do microbiomes cluster by obese and lean mouse phenotypes.

6.6.3 Train SVM to Predict Host Obesity and Host Diet from Microbiome Community Data

The SVM approach for linking a microbiome community population structure and a community metabolome with HMI types has produced excellent results in previous portions of this analysis as well as previously published results (Larsen et al. 2015a, Larsen and Dai 2015). Therefore, we turn to this method for linking mouse microbiome community to predicting obesity in its host.

Metabolic models will be used to generate predictive SVM for two host states: obesity and HF diet. We anticipate that the best predictors of host obesity will not necessarily be predictors of a microbiome’s response to a host’s HF diet.

Models were trained on two datatypes: community structure and metabolomic models. All 20 taxa (**Table 6.3**) will be used in Community Structure SVMs. For training SVM on metabolic model data, all metabolites with unique PRMT-scores were ranked by Fisher score (**Eq. 5.2**). SVMs were trained on all 1386 metabolites and the top Fisher-score ranked 700, 600, 500, 400, 300, 250, 200, 150, 100, 50, and 25 metabolite subsets. SVMs were generated using the same approach as previously described.

Accuracy of predictions were calculated as Mathews Correlation Coefficients (MCC). MCC is calculated as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \qquad \textbf{Eq. 6.8}$$

Where *TP* is the number of true positives, *TN* is the number of true negatives, *FP* is the number of false positives, and *FN* is the number of false negatives. An MCC of 1 indicates a perfect prediction, 0 indicates no prediction, and -1 indicates an inverse prediction.

Results of SVM predictions are summarized in **Figure 6.14**.

From SVM results, community structure data is an excellent predictor of host HF-diet for both training and validation subsets. Metabolic models, although the results from training datasets show strong predictions for a host HF-diet, are not predictive of host diet in the validation dataset subset. For obesity, results are even less encouraging. Although both taxa and metabolism perform well in training sets, neither data type is predictive for validation data. Further, neither taxa nor metabolism is capable of predicting the host's phenotype for 'Transplant' mouse data indicating that SVMs may be suffering from overfitting in these models. So far, our analysis closely concurs with that of Xiao et al. (Xiao et al. 2017): HF-diet strongly drives mouse microbiome community, but host obesity does not leave a clear signal in the microbiome.

These results indicate that an alternative approach for distinguishing the differences between the microbiome's response to host diet and the host's phenotypic changes in response to its microbiome will be required.

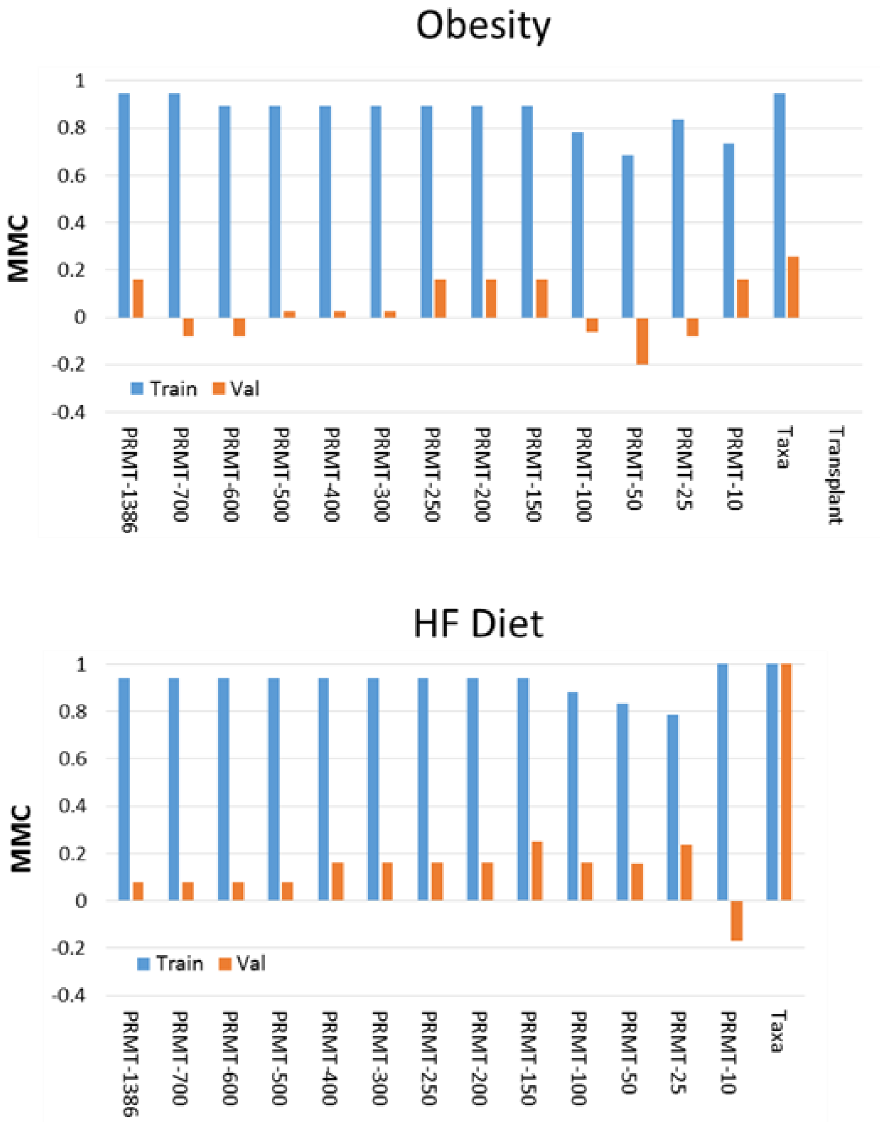


Figure 6.14. Results for SVM prediction of Obesity and HF Diet from microbiome metabolome and community structure. Results are presented as MCC scores for both training (blue bars) and validation (orange bars) sets. The nature of the data used to train the SVM are listed on the X-axis: Top Fisher-score ranked subsets of metabolic model data (PRMT), or microbiome community structure (Taxa). For Obesity HMI, an additional dataset was included: data from 'Transplant' microbiome experiments.

In the next section, a Machine Intelligence (MI) approach, utilizing the Nutonian ‘Eureqa’ tool (http://www.nutonian.com/) for predicting a host phenotype from the microbiome is attempted as an alternative to SVMs.

6.6.4 Machine Intelligence Prediction of Obesity from Microbiome

As an alternative to the SVM approach, the Machine Intelligence (MI) approach will be utilized for construction predictive models of HMI for host obesity and host HF-diet types. For the ‘Eureqa’ MI method, two different datatypes were considered. For the ‘Taxa’ dataset, all 20 taxa were used for model construction. For the ‘Metabolism’ dataset, the 128 PRMT metabolites with the highest Fisher scores (Obese vs Lean microbiomes) were considered in the model. MI equations were set up as the following:

$$\textit{Obesity-score} = f\,(\textit{model features}) \rightarrow \begin{cases} 0 \text{ for non-obese host} \\ 1 \text{ for obese host} \end{cases}$$

and

$$\textit{HFdiet-score} = f\,(\textit{model features}) \rightarrow \begin{cases} 0 \text{ for LF host diet} \\ 1 \text{ for HF host diet} \end{cases}$$

Where ‘*model features’* are the 128 PRMT scores or the 20 taxonomic abundances in microbiomes. Functions were generated in ‘Eureqa’ using the equation elements addition, subtraction, multiplication, division, and constant, and the MI evolutionary equation fitting algorithm was allowed to run until stability and convergence were greater than 95%. The result of these functions is termed an Obesity Score (OS) or HF-diet Score (HFS).

It is expected that although the functions are trained on values of 1 and 0, most applications of the functions will potentially fall into intermediate values. For establishing host phenotypes from OS, we have chosen to use a Relative Propensity for Obesity (**RPO**) score and Relative Expectation for HF Diet (**RHF**) score. The RPO-score and RHF-score are calculated as:

$$\textbf{RPO} = \textbf{log2}(OS\ /\ \textbf{Ave}(\text{all } OSs)\ ) \qquad \textbf{Eq. 6.9}$$

$$\textbf{RHF} = \textbf{log2}(HFS\ /\ \textbf{Ave}(\text{all } HFSs)\ ) \qquad \textbf{Eq. 6.10}$$

Where values less than zero indicates assignment of the phenotype 'Obese' or 'HF-diet' to microbiomes.

The resulting functions for the prediction of host obesity and host HF-diet from microbiome data were:

$$OS_{met} = 2.434 + 0.5053*MET1152 + MET1152/(0.4903 - MET904) - MET1238*MET051\text{^}2 \qquad \textbf{Eq. 6.11}$$

$$OS_{taxa} = 87.68*Por + 0.4109*Bec*Eub + Cls*Col*Lact - Eub - Des*Par \qquad \textbf{Eq. 6.12}$$

$$HFS_{met} = 1.027 + MET070*MET063 + 8.764*MET070*MET1360 + 1.681*MET619*MET070\text{^}2*MET063\text{^}2 - 2.246*MET070*MET1018 \qquad \textbf{Eq. 6.13}$$

$$HFS_{taxa} = (8558102.7*Col*Rumi\text{^}5 - 12.59)/(3784.4 + 8547675.6*Col*Rumi\text{^}5) \qquad \textbf{Eq. 6.14}$$

In Equations 6.11-6.14, the relative abundances of bacterial taxa are identified by the first 3 or 4 unique letters of the full taxonomic name. Metabolism MET# identifiers are associated with their metabolic model compound in **Table 6.6**. The results of MI models predictions of obesity and HF-diets are summarized in **Figure 6.15**.

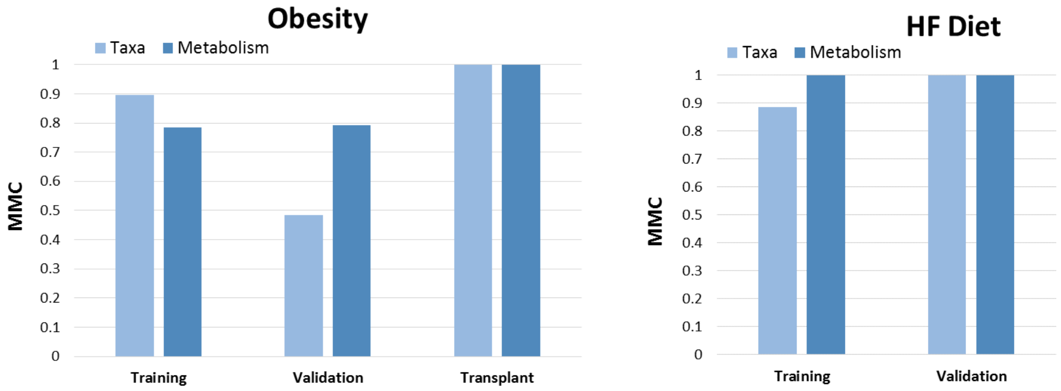


Figure 6.15. Results for MI prediction of Obesity and HF Diet from microbiome metabolome and community structure. Results are presented as MCC scores for both training (blue bars) and validation (orange bars) sets. The nature of the data used to train the SVM are listed on the X-axis: Top Fisher-score ranked subsets of metabolic model data (PRMT), or microbiome community structure (Taxa). For Obesity HMI, an additional dataset was included: data from 'Transplant' microbiome experiments.

Results from the MI model for prediction of a HF-diet show that both taxa and metabolism are good predictors of host diet, with MCC validation results of 1.0 for both data types. However, for prediction of host Obesity, metabolomic data is found to be a stronger predictor than taxa (MCC of 0.79 and 0.48 respectively). This is in accordance with our previous results (Larsen et al. 2014, Larsen and Dai 2015) that indicate that a microbiome community metabolism is a better predictor of HMI than community structure. MI models are also far better predictors of both HF-diet and Obesity than SVM trained on the same data. Most importantly, accurate predictions of obesity of the 'Transplant' dataset indicates that the model generated here may be generalized beyond the 'Obesity' dataset and will be a useful tool for predicting the interactions between diet, microbiome, and host obesity in other microbiome datasets.

Table 6.6. Metabolites predictive of host HF-diet and obesity in MI models

	ID	Metabolite	KEGG Pathway
HF-Diet	MET063	N1-(5-Phospho-alpha-D-ribosyl)-5,6-dimethylbenzimidazole	Metabolism of cofactors and vitamins
	MET070	Phenanthrene-4,5-dicarboxylate	Polycyclic aromatic hydrocarbon degradation
	MET1018	Phosphonoacetaldehyde	Amino acid metabolism
	MET1360	L-Rhamnulose 1-phosphate	Carbohydrate metabolism
	MET619	1D-1-Guanidino-3-amino-1,3-dideoxy-scyllo-inositol	Antibiotic biosynthesis
Obesity	MET051	1-Acylglycerol	Glycerolipid metabolism, Fat digestion and absorption, Vitamin digestion and absorption
	MET1152	alpha-Oxo-benzeneacetic acid and 4-Hydroxyphenylglyoxylate	Phenylalanine metabolism
	MET1238	Deoxycytidine	Pyrimidine metabolism
	MET904	Leukotriene D4	Neuroactive ligand-receptor interaction, Arachidonic acid metabolism, Fc epsilon RI signaling pathway

The metabolites identified as predictive for host HF-diet and obesity also provide insights into molecular mechanisms of HMIs. The difference between HF and LF diet in a microbiome community structure can be defined exclusively by the relative abundances of *Collinsella* and *Ruminococcaceae*. Metabolic functions associated with a HF diet are primarily associated with bacterial metabolism: amino acid metabolism, carbohydrate metabolism, biosynthesis of co-factors, and metabolism of complex ringed molecules. This is consistent with a microbial population that changes its community structure in response to new nutrient sources, which in this case is the different sugar and fat contents between a LF and HF host diet. Those predicted microbiome features that are associated with host obesity however are very different in nature. Seven of the twenty taxa in the microbiome community structure are associated with host obesity in the MI model. *Bacteroides*, *Clostridium*, *Lactobacillus*, and *Parabacteroides* have previously been associated with obesity (Million et al. 2012, Le Chatelier et al. 2013, Leung et al. 2013, Walsh et al. 2014, Kasai et al. 2015). There is no direct association between *Eubacterium* and obesity found in the available literature, but the predicted mouse gut microbiome CIN and dynamic MAP-model from the previous section in this analysis, the abundance of *Eubacterium* is positively associated with dietary fat intake. The molecules and functions associated with obesity are primarily associated with host interaction. "Glycerolipid metabolism, Fat digestion and absorption, and Vitamin digestion and absorption" are pathways associated with the host's ability to absorb nutrients from diet rather than a bacteria's capacity to consume them. Pathways "Neuroactive ligand-receptor interaction, Arachidonic acid metabolism, and Fc epsilon RI signaling pathway" seem to point directly to the specific molecules that mediate interactions in the gut-brain axis, interfacing the microbiome community directly with the host's regulatory networks and perhaps even the host's behavior. Leukotierenes are directly associated with obesity (Back et al. 2014), inflammatory pathways (Busse 1998), and response to insulin (Martinez-Clemente et al. 2011, Li et al. 2015). Phenylalanine pathways have been previously observed to be highly enriched in the microbiomes of obese hosts (Liu et al. 2017) and pyrimidine metabolism has been observed to be reduced in non-obese animals (Yang et al. 2016a). 4-Hydroxyphenylglyoxylate is an inhibitor of fatty acid oxidation that can lead to liver disease and affect the digestion of fatty acids in the gut (Keung et al. 2013). Cytidine deaminase, the enzyme responsible for deoxycytidine metabolism in the obesity-predictive metabolites is also known to be linked to obesity-associated reduction of immune B-cell responses (Frasca et al. 2008, Frasca et al. 2016).

6.6.5 Summary of Results

In this Task, our goal was to generate predictive models of host obesity from microbiome data. A unique, previously published dataset using mouse genetic variants was used to train models. This dataset allowed models to distinguish between the effects of HF-diets on the microbiome and the effects of the microbiome on the host's obesity. Unlike in previous results, both in this study and in our previously published manuscripts (Larsen et al. 2015a, Larsen and Dai 2015), using SVM to predict HMI from mouse microbiome data was found to suffer from overfitting. While microbiome community structure was found to be a good predictor of host HF-diet by SVM, for all other conditions and data types, the validation of SVMs were weak. An alternative approach for predicting obesity from microbiome data needed to be found.

Using an MI approach, specific functions were found that could accurately predict both host obesity and host HF-diet from microbiome data. Here, as in our previous analyses, we find that emergent properties of the microbiome community, specifically the emergent property of the microbiome community metabolome, are more predictive of HMI than is microbiome community structure alone. More importantly for our proposed research plan, the MI approach, when trained on 'Obesity' data could predict obesity in microbiomes observed in the 'Transplant' dataset. This ability to extrapolate predictions to very different microbiome population structures across different experimental observations validates our approach and indicates that this will be an effective tool for predicting HMIs for multiple biological conditions.

Predictions of host obesity and host HF-diets also provide insights into some of the specific mechanisms of HMIs. The microbiome's responses to a change in the host's diet were exclusively metabolic in nature: changing metabolic capacity to accommodate the change in available nutrients from a LF to HF diet. The microbiome's influence of host obesity however, was entirely associated with synthesis of molecules that interact with the host's immune system, induce inflammation in the host, or signaling mechanisms for the gut-brain axis. Obesity in this model then is not primarily about the microbiome changing the amount of energy that can be extracted from the host's diet, but rather due to the microbiome's influence on a variety of host immune responses.

6.7 Combine Computational Models for System-scale Predictions of Host-Microbiome Interactions

A suite of microbiome-related models have been proposed and validated in prior Tasks in this Chapter:

- Dynamic MAP-model predicts changes in a microbiome community resulting from the host's diet.
- TAvP-prediction extrapolates EFPs from a microbiome community structure using a statistical analysis of thousands of previously sequenced and annotated genomes.
- PRMT can be used to generate microbiome community metabolic models from EFPs.
- RPO-scores, for predicting host obesity from the microbiome community metabolome, have been constructed from community metabolomic model data.

All analysis engines specific developed in this analysis have been constructed and validated using a variety of prior biological observations, with each model component drawing from different available mouse microbiome datasets (**Figure 6.16**).

Here, we combine analysis engines to create iMOUSE, a dynamic, system-scale model of host-microbiome interactions for prediction of host obesity. The complete iMOUSE model is validated by repeating published biological experiments *in silico*. The validated system-scale model will then applied to a novel biological question: mice with an Obese microbiome transplant have an increased propensity for obesity even when on a LF-diet, but is there a configuration of diet parameters that can produce a 'Lean' host phenotype in a mouse with an 'Obese' microbiome?

This Task will accomplish several goals. The first is to combine previously generated and validated computational models of the mouse microbiome into a single, system-scale model. The second is to validate the iMOUSE model by reproducing, *in silico*, the ‘Gradient’ biological experiment and to apply the ‘Gradient’ experimental design to the ‘Transplant’ ‘Lean’ and ‘Obese’ microbiome transplant community structures. Finally, a Genetic Algorithm (GA) approach will be applied to the iMOUSE model to identify diets that will provide a ‘Lean’ phenotype for mice with an Obese-transplant microbiome.

6.7.1 The iMOUSE Model

The specific mechanisms used to link previous models into the Mouse HMI System Model, iMOUSE, are shown in **Figure 6.16**. Some modifications however needed to be made to model components. In previous Tasks in this chapter, methods for predicting an EFP of thousands of enzyme functions and metabolic models with thousands of metabolites have been demonstrated. However, only a very small fraction of those metabolites were identified as relevant for the prediction of host obesity (**Table 6.6**). Therefore, we propose the concept of the Obesity-Specific Metabolome, which includes only those metabolites needed for prediction of host obesity and those enzyme functions required for modeling those metabolites for the system-scale model of mouse obesity HMIs (**Table 6.7**).

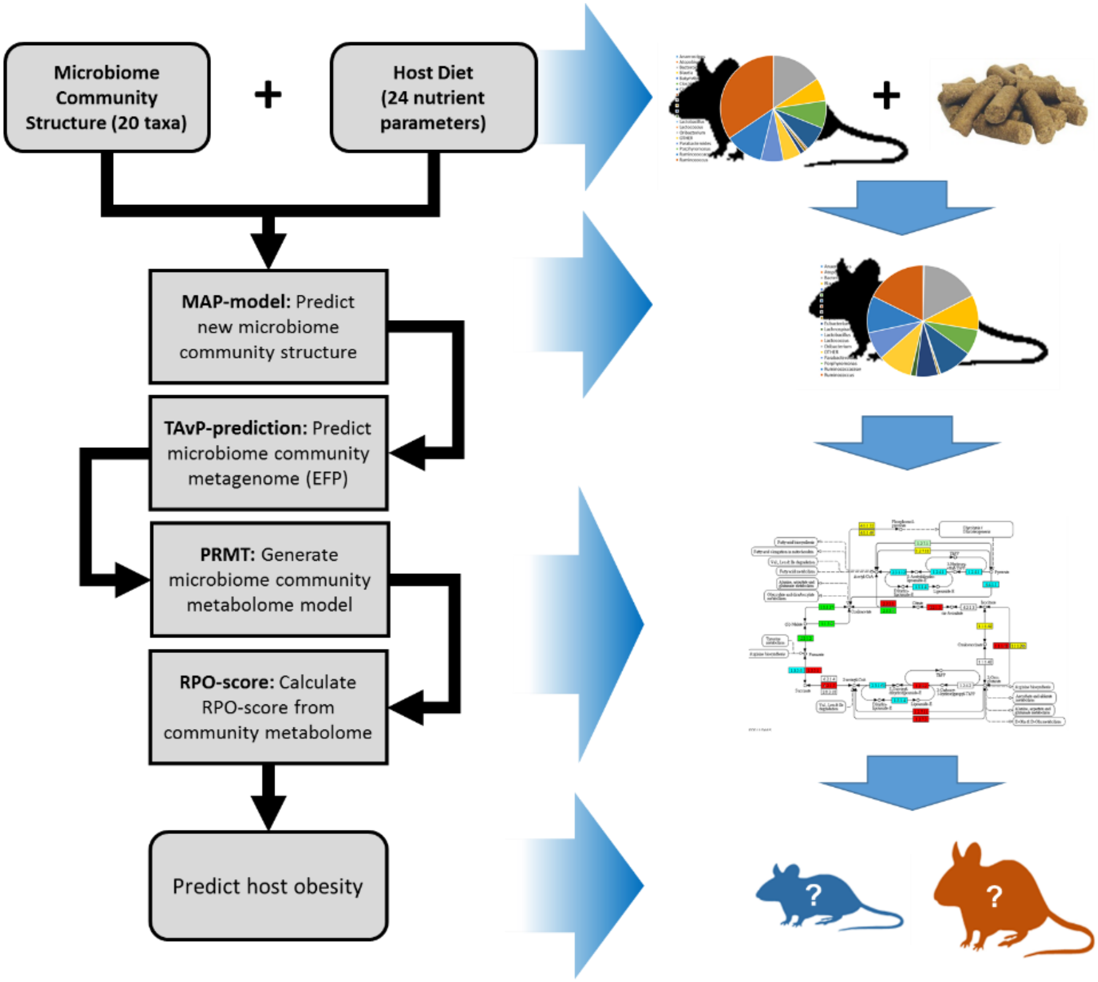

Microbiome Community Structure (20 taxa)
+
Host Diet (24 nutrient parameters)
MAP-model: Predict new microbiome community structure
TAvP-prediction: Predict microbiome community metagenome (EFP)
PRMT: Generate microbiome community metabolome model
RPO-score: Calculate RPO-score from community metabolome
Predict host obesity
+
?
?

Figure 6.16. iMOUSE, a system-scale model of mouse HMI. By combining models created and validated from a wide variety of microbiome datasets and databases of annotated sequences genomes, a complete system-scale model of mouse obesity HMI can be constructed. In this model, starting with an initial microbiome community structure and a set of diet parameters, a dynamic MAP-model will be used to determine how the microbiome community changes in response to diet conditions. From predicted microbiome community, an Obesity-Specific EFP is created and used to calculate the Obesity-Specific metabolome. The Obesity-Specific metabolome is used to calculate an RPO-score for the prediction of a host's predisposition for obesity

Table 6.7. The Obesity-specific HMI metabolome

Obesity-Related Metabolome		Function
Enzyme Functions	2.3.2.-	Gamma-glutamylaminecyclotransferase
	2.3.2.2	Gamma-glutamyltransferase
	3.1.1.23	Acylglycerol lipase
	3.1.1.3	Triacylglycerol lipase
	3.1.3.5	5'-nucleotidase
	3.5.4.5	Cytidine deaminase
	4.1.1.7	Benzoylformate decarboxylase
Metabolites	1-Acylglycerol	Glycerolipid metabolism, Fat digestion and absorption, Vitamin digestion and absorption
	alpha-Oxo-benzeneacetic_acid	Phenylalanine metabolism
	Deoxycytidine	Pyrimidine metabolism
	Leukotriene_D4	Neuroactive ligand-receptor interaction, Arachidonic acid metabolism, Fc epsilon RI signaling pathway
	4-Hydroxyphenyl-glyoxylate	Biosynthesis of antibiotics

6.7.2 Validate iMOUSE Model by reproducing prior biological experimental results in silico

The iMOUSE model outlined in **Figure 6.10** will be validated using two host-microbiome *in silico* experiments:

1) ***In Silico* 'Gradient' Experiment.** Here, we use iMOUSE model to recreate the 'Gradient' experimental data. From a predicted microbiome structure, generate a community metabolome model and predict the relative propensity for obesity that derives from the 'Gradient' diets, ranging from 100% HF to 100% HF diets.

2) **Predict effect of diet on 'Obese' and 'Lean' microbiome transplants for mouse obesity HMI.** We hypothesize that iMOUSE model will predict that mice with an 'Obese' microbiome transplant have greater propensity to obesity than mice with a 'Lean' microbiome transplant, matching the experimental observations.

Each experiment can be described using the following conditions. The starting microbiome community structures used in the *in silico* experiment and the set of diet conditions that are applied to the starting microbiomes. The results from the iMOUSE model are predicted microbiome community structures, predicted obesity-related EFPs and metabolomes, and calculated RPO-scores.

6.7.2.1. In Silico *Diet Gradient Experiment*

Goal: The goal of this experiment is to reproduce *in silico* the results of biological experiment that generated the 'Gradient' dataset.

Starting microbiome conditions: From the 'Gradient' dataset, microbiome community structures from the five 'control' microbiomes were selected that, in the initial experiment, were maintained on a LF diet both for Initial and Final microbiome conditions.

Diet Conditions: The same diet conditions – 0, 1, 10, 25, 50, 25, and 100% HF-diet – as used in the initial 'Gradient' experiment were used in this *in silico* experiment.

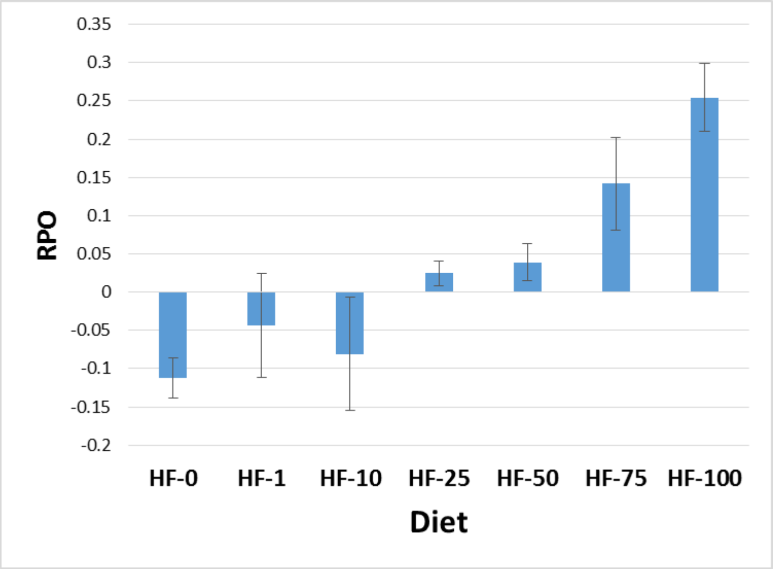


Figure 6.17. Predicted Relative Propensity for Obesity (RPO) scores for mice on increasingly HF diets. Average and standard deviations for predictions generated from replicated experimental conditions. In these models, a higher RPO-score indicated a greater propensity for obesity in the host. Diets are comprised of a mix of HF and LF pellets, where HF-*n* indicates a diet that is composed of *n*% HF pellets.

Results: The resulting RPO-scores are summarized in **Figure 6.17**. With increasing proportion of HF diet, there is a corresponding increase in RPO-scores.

Conclusions: Results closely follow biological expectations: with increasingly HF-diets, the resulting HMI's are increasingly predicted to produce an 'Obese' phenotype in the host. What makes these results significant is that the model correctly extrapolates obesity in 'Gradient' data mice when 'Gradient' data was not used in training the Obesity function. Furthermore, community structures from 'Gradient' and 'Obesity' are quite distinct from one another, suggesting that Obesity Function has indeed identified a potential underlying biological mechanism linking the microbiome community metabolism to obesity HMI.

6.7.2.2. In Silico *Microbiome Transplant Experiment*

Goal: This *in silico* experiment is more ambitious than the previous. First, this experiment utilizes data not used to train any component of the system-scale model. Second, this experiment attempts to reproduce a biological phenomenon that did not appear in any of the model training data, specifically that an Obese-microbiome transplant predisposes the host to obesity regardless of diet.

Starting microbiome conditions: This experiment uses the 'Lean' and 'Obese' microbiome community transplant structures from the 'Transplant' dataset.

Diet Conditions: The same diet regimes used in the 'Gradient' experiment (i.e. 0, 1, 10, 25, 50, 75, and 100% HF diets) are used here.

Results: RPO-scores are presented in **Figure 6.18**. Here it is shown that mice with an Obese microbiome transplants are constantly more likely to have an obese phenotype regardless of diet conditions. Obesity for both 'Lean' and 'Obese' microbiome transplants is increased by a HF-diet.

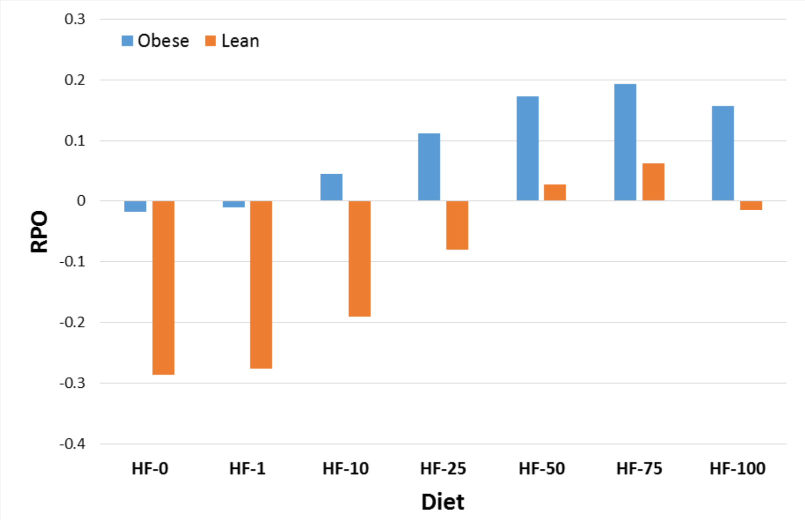


Figure 6.18. Mice with 'Obese' microbiome transplants have increased predicted propensity for obesity over mice with 'Lean' microbiome transplants. While increasing HF-content of diet increases the likelihood for obesity in mice with both 'Obese' and 'Lean' microbiome transplants, mice with 'Obese' microbiome transplants are predicted to be far more prone to obesity at all diet types. Diets are comprised of a mix of HF and LF pellets, where HF-*n* indicates a diet that is composed of *n*% HF pellets.

Conclusions: Results are strongly supportive of the hypothesis that we can model the effects an 'Obese' microbiome that predisposes the host to an obese phenotype. In spite of not being trained on the 'Transplant' data, the model correctly predicts several key biological observations. An 'Obese' microbiome is more likely than a 'Lean' microbiome to result in an obese mouse, regardless of diet.

6.7.2.3. Summary of iMOUSE in Silico *Experiments*

These results demonstrate the robust extrapolative power of the iMOUSE model. The 'Transplant' dataset was not used to train any portion of the component models in the system-scale model. Not only does the system-model correctly reproduce in silico the results of the 'Transplant' biological experiment, but the model also correctly extrapolated the diet-resistant effects of an 'Obese' microbiome in spite of the fact that there were no equivalent observations in the data used to train the models. These results are very strongly suggestive that the *in silico* system-scale model of mouse HMIs has accurately captured relevant biological interactions capable of making relevant predictions for the outcome of biological experiments.

6.7.3 Use Genetic Algorithm for Identifying 'Diet for Lean Phenotype

The Mouse HMI System Model has been effectively validated through the successful prediction of multiple biological observations from experiments not explicitly used to train the models. Here, our goal is to apply the validated system model to determine a potential diet that restores 'Lean' phenotype to mice with an 'Obese' microbiome transplant. Our approach is to use a GA approach to search the range of possible diet parameter combinations that minimizes RPO scores for modeled microbiomes that have initial community compositions from 'Lean' and 'Obese' transplant microbiomes. We anticipate three possible outcomes to this approach: (i) restoration of lean microbiome community structure from an initial obese microbiome community, (ii) identification of diet that will restore lean phenotype in spite of obese microbiome, or (iii) identification of a novel microbiome state achievable from starting obese microbiome, that will yield a lean phenotype.

One possible approach would be to screen all possible diet parameter combinations and select the one which minimized OS. This however would necessitate the calculation of around $2x10^{26}$ possible diets (24 diet parameters, each with a possible integer value from 1 to 100, for two possible starting microbiome community structures). Even if each diet condition could be evaluated in one hundred thousandths of a second, it would still take about $6.34x10^{11}$ processor years to search this space. We utilize a variation of a Genetic Algorithm (GA) approach to somewhat reduce computation time.

In a GA approach, populations of possible solutions are randomly generated. The best solutions from every ‘generation’ are selected, then used to make the next population of novel solutions by randomly mixing and mutating solutions in the previous generation of solutions. The GA approach described below follows a ‘bacterial population sharing an antibiotic resistance phenotype’ biological model rather than a ‘genetic hybrid offspring of two parents’ model for its evolutionary process (Overballe-Petersen et al. 2013).

The GA approach utilized the following inheritance rules:

- The previous generation of solutions are sorted from best to worst by Obesity Score
- The best/highest ranked diet always survives unchanged to the next generation
- One copy of the best diet, subjected to random mutations always survives to the next generation
- For every diet after that, each diet parameter has a chance of being replaced by the parameter of a higher ranked diet
    - The probability of a diet parameter being replaced increases with diet rank. Weights are such that the lowest ranked diets are, on average, almost completely derived of parameters swapped from higher-ranked diets

  - Selection of the diet from which the swapped diet parameter is derived is weighted such that higher-ranked diets are more likely to be the source of the cross-over

- Each diet after the highest ranked diet is mutated in 0 to *max_mutation* parameters

This approach is implemented in the method described by the following pseudocode:

```
Pseudocode for Optimizing Diets for any Initial Microbiome Community
Structure

Given:
Microbiome = Starting Microbiome Populations
Initial Diet conditions = vector of diet parameters

maxPop = total number of individuals in population
mut = mutation frequency
gen = Number of generations

Functions:
SortResults: sort population of diets by ObesityFunction values
iMOUSE: Given a diet condition and starting population structure,
           -  calculate new microbiome community structure,
           -  calculate obesity-related EFP,
           -  calculate obesity-related PRMT,
           -  return Obesity Function score calculated from
              metabolome

GA_Diet(Microbiome)
    For g = 1 To gen

        Call SortResults

       #### Cross breed current diets…

        For diet = maxPop To 2 Step -1
            For dp = 1 To 24 #-- there are 24 diet parameters…
                If Rnd * maxPop + 1 < diet Then
                    c = 1
                    Do While Rnd < 0.5 And c < diet - 1
                        c = c + 1
                    Swap diet parameter dp with value from $c^{th}$-ranked
                       diet
                End If

        Save the highest ranked diet unchanged
        #--- Add mutations…
        For diet = 2 To maxPop
            m = Int(Rnd * mut)
            If m > 0 Then
                For i = 1 To m
                    dp = Int(Rnd * 24) + 1
                    nv = Rnd * 94 + 5
                    diet parameter dp in $diet^{th}$ ranked diet = nv
                Next i
            End If
        Next diet

        For each diet in population
           iMOUSE(community, diet)

    Next g

    #--- a final sort to make sure best diet is identified
    Call SortResults

-END
```

The GA optimization method was run using ‘Lean’ and ‘Obese’ microbiome transplant community structures for initial population conditions and LF diet for initial diet parameters. GA was run with a population size of 50 diet parameter combinations for 100,000 generations. Through observations of method implementation, the GA algorithm reached a ‘best’ solutions after approximately 30,000 generations with no further decrease in Obesity Function observed after that.

6.7.3.1 Optimized Diets are Different Depending on Starting Microbiome Community Structure

Diet parameters that result in the most Lean phenotypes, by Obesity Function-score are different depending on whether the mouse starts with a ‘Lean’ or ‘Obese’ transplanted microbiome (**Figure 6.19**).

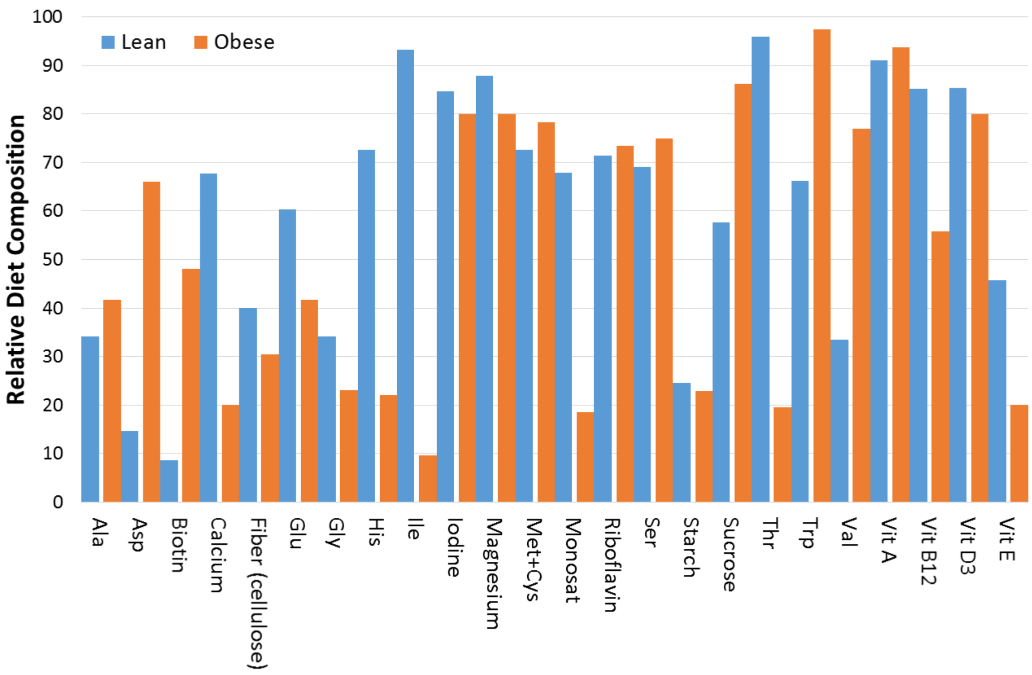


Figure 6.19. Optimized diet parameters for achieving a 'Lean' HMI phenotype from a starting 'Lean' or 'Obese' microbiome community

6.7.3.2 Optimized Microbiome Community Structures are Different Depending on Starting Microbiome Community Structure

The final population structures that result from optimize diet conditions are very different depending on whether the mouse began with a 'Lean' or 'Obese' microbiome transplant community structure (**Figure 6.20**).

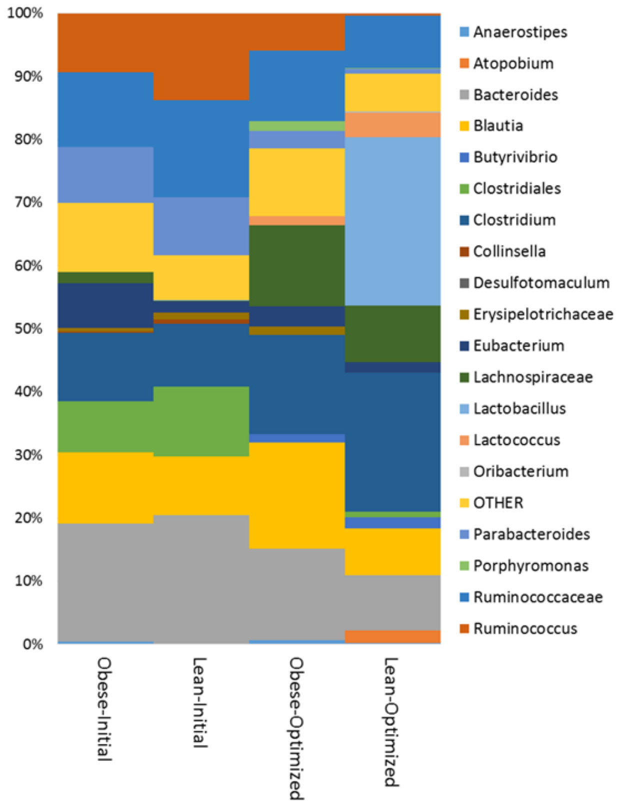


Figure 6.20. Initial and final microbiome community structures in diet optimizations. Population structures are for the initial microbiome 'Obese' and 'Lean' transplanted communities and for the community structures that result from diets that optimize microbiome community RPO-scores.

6.7.3.3 Optimized EFPS are Very Similar, Independent of Starting Microbiome Community Structure

These results indicate that very different population structures can have very similar enzyme profiles and, therefore, similar effects on the host-microbiome interaction (**Figure 6.21**). This agrees with our initially stated hypothesis that HMIs are an emergent property of a microbiome community.

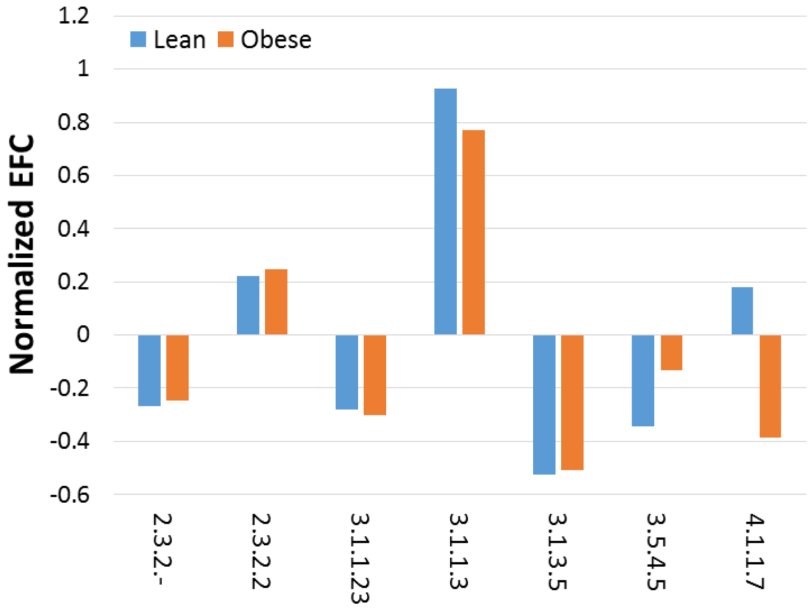


Figure 6.21. Obesity-specific EFPs for diet-optimized Lean and Obese microbiome communities.

6.7.4 Validation of iMOUSE Model

Without follow up biological experiments, there is no way to truly validate how accurate are the predictions for optimal diet parameters targeting ‘Lean’ and ‘Obese’ microbiome community structures. We can, however, consider the results in light of prior biological knowledge to gauge how ‘reasonable’ the predicted solutions are through a consideration of what is previously known about the distributions of mouse microbiome community structures.

6.7.4.1 Are diets ‘reasonable’?

It may be that the GA approach is capable of generating a diet that nurtures the microbiome community but leaves the host to starve. We can ask of the results, how well do the optimized diets fall within parameters of known diets. In model development stages, all diet parameters were normalized to arbitrary values between 20 and 80. Optimized diets were restricted to values between 1 and 100 for all diet parameters, so no predicted diet can be more than 20 arbitrary units outside of the actual observed range of diet parameters. In the diet optimized for 'Obese' microbiomes, there are 3 parameters below 20 units and 7 parameters above. For the diet optimized for 'Lean' microbiomes, there are three diet parameters each above 80 or below 20. The majority of diet parameters then fall into intermediate levels between standard HF and LF-diets including the parameters for fats, sugars, starch, and fiber. Perhaps unsurprisingly, the optimal diet for 'Obese' microbiomes is slightly reduced for saturated fat content (18.6 arbitrary units) relative to standard laboratory diets. Overall, there seems to be no reason to believe that predicted diets are incompatible with supporting the host mouse. The answer to our initial question then is 'yes', the predicted diets are reasonable ones.

6.7.4.2 Are final population structures 'reasonable'?

It is also possible that the predicted microbiome populations, while predictive of a Lean HMI phenotype, are biologically implausible. Again, while there is no way outside of additional biological experimentation to determine if the predicted population structures are actually achievable, we can turn to the existing observation of microbiome community structures to see if the predicted structures are likely. In the predicted community structures, there are no taxonomic abundances that exceed those in the collected microbiome populations ('Oscillation', 'Catalog', 'Obesity', and 'Transplant') except *Blauta* in the Obesity microbiome optimized data. *Blauta* are associated with healthy gut microbiomes and decreased host inflammation (Hong et al. 2011, Bajaj et al. 2012) . The abundance of *Lactobacillis* in the Lean microbiome optimized community structure is high, but not higher than has been previously reported in the set of communities analyzed here. The lean microbiome-optimized diet correlated most closely (by PCC between $\log_2$ transformed relative population abundances) with a 'Catalog' microbiome for a male SV129 mouse on LF diet. The obese-microbiome diet correlated most strongly with a 'Catalog' microbiome for a male C57BL/6 mouse on a LF diet. So again, the answer to our question is ‘yes’, there is no reason to consider that the microbiome community structures that result from diet optimization to be unrealistic.

7. CONCLUSIONS

Microbiomes play a crucial role in living systems and have profound and diverse effects on their hosts. Understanding the molecular mechanisms of HMI are complicated however by the tremendous diversity of microbiome communities. Here, through application of our central hypothesis that HMI are an emergent property of the microbiome community, we have investigated HMI's at multiple scales and in multiple biological systems. These studies have resulted in the creation of iMOUSE, a system-scale model of mouse obesity HMI. Through these analyses we have accomplished two principle goals. The first is to unambiguously support our central hypothesis that HMI is an emergent property of microbiomes. The second is to generate a set of computational models of HMI that span multiple biological scales, accurately replicate previously reported biological experiments *in silico*, and can propose novel biological approaches for the rational manipulation of microbiome communities for the benefits of the host's health.

7.1. Pseudomonas-Host Interactions

Genomic information and models built from annotated genomes were used to accurately predict plant root and human host HMI classes in Pseudomonads. Emergent properties of the genome, the metabolome and the transportome, are more predictive than simply the identity of the genes in a genome. In all cases but one, genomic information alone (in the format of Enzyme Function Profiles) are the least accurate predictors of HMI classes. In every case but one (Plant Disease), transportomic models (i.e. PRTT-scores) is the most predictive data type. This suggests that Transportome, the capacity of a microorganism to sense, manipulate, and engage directly with its environment, is the most biologically relevant capacity for a Pseudomonad to interact with a host organism. These findings support the hypothesis that HMI derive from the emergent properties of an interacting network of genes and proteins that dictates a Pseudomonad's ability to interact with its environment and there are multiple possible arrangements of genes that are able to achieve that HMI class.

While the predictive abilities of this approach are strong, the potentially far more biologically relevant results are insights into the specific molecular mechanisms that enable human and plant pseudomonad-host interactions. Those mechanisms, with regard to human pathogenicity, point to possible methods for disrupting antibiotic resistance (e.g. quorum sensing, biofilm formation, or bacteriocin production) and interfering with the capacity of a pathogen to evade the host immune system (e.g. arginine and dopamine biosynthesis). These molecular targets may lead to new human therapeutic interventions that address the rising dangers of antibiotic resistant bacterial infections. With regard to enhancing the potentially positive PGP activities of Pseudomonads, the predicted molecular mechanisms of host interactions provide insights as well. Possible mechanisms to increase the PGP effects in Pseudomonads include manipulation of biosynthetic pathways for plant-signaling compounds such as auxin (indole) or neringenin. Greater benefit to a plant host by Pseudomonads might be achieved by enhancing bacterial production of antimicrobial or antifungal compounds. These hypotheses, generated from computational analysis of HMI and available genomic data, are directly testable by laboratory experiments.

The method proposed here of using metabolomic and transportomic models to classify Pseudomonas by HMI class can be generalized to other HMI classes and other species of bacteria. There is no reason that a far wider range of organisms than Pseudomonads could not be considered, and the list of possible HMI classes for analysis is vast. We expect that this analysis approach will be a useful and powerful tool for genomic analysis for a wide range of bacteria and bacterial-host or bacteria-environmental interactions.

7.1.1. Potential Weaknesses

One possible weakness of this study is the relatively small set of available genomes suitable for analysis. While there are hundreds of possible sequenced Pseudomonas genomes available in NCBI and thousands of genomes or partial genomes collected from metagenomics studies, very few of those can also be confidently traced to a specific set of HMI classes by direct observation in the reported literature or through experimental laboratory manipulations. We fully expect that, as additional organisms are sequenced and characterized, the method described here will successfully incorporate them into analyses and the predictive power of this approach will be further improved.

Also, the definitions and assignation of HMI classes to Pseudomonads may not be optimal. HMI classes, gathered from Silby et al. (Silby et al. 2011), and verified by investigation into the primary literature, still may be too coarse grained for optimal descriptions of different bacteria-host interactions. Here, a necessary balance between tractable numbers of HMI classes that can be ascribed to a reasonable number of individual species needed to be selected. In future work, additional HMI classes of greater specificity with a larger number of representative species for use in training computational models will be required.

Fortunately, these potential weaknesses can be easily addressed with the addition of more and more Pseudomonas species for which HMI classes can be determined. With the vast and rapid increase in sequenced genomes being deposited in public databases there will be an ever increasing resource of well-characterized genomes for use in analysis.

7.1.2. Proposed Future Work

The results of this analysis show that much additional research can still be accomplished. The approach presented here was highly accurate at predicting HMI types from genomic model data for Pseudomonads, and we anticipate that this approach will be a genome annotation and analysis tool that has great value to the scientific community, particularly for the annotation of newly sequenced bacterial genomes assembled directly from metagenomic sequence data. For this to happen, a more accessible analysis pipeline must be created using the principles we have shown. Currently, we are working with researchers at KBASE (http://kbase.us/) to establish our tool as a standard analysis pipeline that will automate many of the genome annotation and SVM training and validation steps demonstrated here.

KBASE is a data analysis platform for systems biology analysis maintained by the Department of Energy. Establishing our analysis method for predicting HMI types from genomic sequence data will ensure that this becomes a common tool for genome analysis.

The results of this analysis had led to a number of specific biological hypotheses that can potentially be experimentally validated. Analysis of pathogenicity in Pseudomonas and the identification of potential molecular mechanisms by which pathogenicity could be mediated would require studies in animal models. It is far easier to design and conduct laboratory experiments for plant-microbe interaction mechanisms than for mammalian-gut microbiome interactions. One such set of experiments is currently underway and has led to published results, described in greater detail below.

7.1.3. Biological Validation: Pseudomonas fluorescens Transportome Is Linked to Strain-Specific Plant Growth Promotion in Aspen Seedlings under Nutrient Stress

In our recent publication (Shinde et al. 2017), we have had an opportunity to experimentally validate the results of the Plant Growth Promotion HMI computational modeling presented here. In this laboratory experiment, aspen (*Populus tremuloides*) seedlings were co-cultured with four Pseudomonas species: one *Pseudomonas protegens* (Pf-5) and three *Pseudomonas fluorescens* strains (Pf0-1, SBW25, and WH6). Cultures were grown under replete, low nitrogen, and low phosphorus conditions. A total of 16 phenotypic measurements were collected from aspen seedlings: dry weight (mg), shoot length (mm), number of leaves, leaf chlorophyll (Chl) concentration (μg mg^{-1} FW), Chl a/b ratio, shoot anthocyanin concentration (μg mg^{-1} FW), shoot NO^{-3} concentration (mg g^{-1} DW), shoot P concentration (mg g^{-1} DW), root dry weight (mg), root branching (integer value), root length (cm), number of rootlets, root anthocyanin concentration (μg mg^{-1} FW), root total N (%), root NO^{-3} concentration (μg g^{-1} DW), and root P concentration (mg g^{-1} DW). The relative capacity of different Pseudomonas species to promote plant growth and mediate nutrient-related stress in aspen seedlings was correlated with Pseudomonas transportomic capacity (PRTT-scores) and results analyzed in the context of the results of this study. Of particular relevance to this current study was the implication of arabinose transport, a plant-stress singling compound, as important to Pseudomonas-plant root interactions for promotion of PGP activities. This experimental observation confirms some of the computational results presented here. In the laboratory, Pseudomonas arabinose transporters have been cloned, synthesized and molecularly characterized. Arabinose transporter knock-out mutants have since been generated in Pseudomonads. At the time of this writing, the arabinose knock-out mutants are being co-cultured with aspen seedlings to validate the role of Pseudomonas arabinose transport in aspen seedling stress remediation and PGP. Thus, a complete cycle of analysis, hypothesis, and experimental validation using the computational approach presented here has been accomplished and will contribute significantly to the understanding of Pseudomonad PGP HMIs.

7.2. Modeling Human Dysbiosis

A longitudinal human microbiome dataset was used to generate and validate computational models for prediction of host dysbiosis, modeling microbiome community dynamics, and predicting diet conditions that will lead to diet-induced dysbiosis.

Dysbiosis is most accurately predicted by the microbiome metabolome, supporting the central hypothesis that HMIs are emergent properties of the microbiome community and less dependent upon the specific presence, absence, or relative abundance of any specific bacterial species or taxa. Model results propose specific molecular mechanisms of dysbiosis, including disruption of fat and protein digestion by the host, disruption of vitamin biosynthesis by the microbiome community, and the presence of potential virulence factors. These predicted mechanisms of host dysbiosis are supported by previously-published experimental observations.

A model of microbiome community dynamics was generated using the dynamic MAP-model approach, which is based on significant technical advancement of our previously published MAP-model method. The dynamic MAP-model predicts that the diet parameter that has the greatest effect of microbiome dynamics is fiber, followed by saturated fats and, potentially, sodium-containing food preservatives. These predictions are corroborated by prior biological knowledge from published literature.

The prediction of diet-induced dysbiosis was accomplished by linking together the model of microbiome community dynamics with the prediction of host dysbiosis from the emergent properties of a microbiome community. Using this novel computational approach, it is predicted that the microbiome community that results after fourteen days of a very high fat diet, a very low fiber diet, or a very low sugar diet induces dysbiosis in the host. These conditions for diet-induced dysbiosis are supported by evidence in the scientific literature.

The power of these results lies within the models capacity to extrapolate well beyond the information used to construct the computational models to propose novel results and generate new biological hypotheses. The model used to identify dysbiosis was not trained on any information derived from diet-induced dysbiosis. The model that predicted microbiome community dynamics did not incorporate any information regarding host dysbiosis. Yet when these models are combined, the resultant system-scale model linking host, host diet, and microbiome community is capable of accurately deducing the dietary conditions – high fat, low fiber, and low sugar – known to induce host dysbiosis. This result indicated that the computational models, far from only being able to return the data on which it was trained, have captured some distillations of biological truth and that these models are valuable engines for understanding HMIs and proposing relevant hypotheses to drive future experiments.

7.2.1. Potential Weaknesses

There are a few key weaknesses in this analysis that should be highlighted. One is the very limited number of people (n=2) in this analysis. The small sample size in this study is typical of microbiome analyses from human subjects, where often many individuals' microbiomes are sampled once or a few times, or a small number of individuals' microbiomes are sampled more frequently. While the depth of data collected in the David et al. experiment is exceptional in its coverage over time, the small number of donors available in the study makes it difficult to determine how likely model results are generalizable outside of the human donors involved in this study. For now, this is a necessary trade off due to the monetary and computational costs associated with microbiome sample collection, sequencing, and analysis. The other main weakness of this analysis is the availability and identification of dysbiotic microbiomes. Donor A's dysbiosis was reported as being due to international travel and Donor B's dysbiosis was due to a specific bacterial incursion by food poisoning. It is possible that while "dysbiosis" accurately describes the donors' phenotypes, two very different mechanisms of HMI are responsible for dysbiosis that arises from these examples of very different dysbiosis- inciting incidents for each donor. This objection is tempered somewhat by the fact that common predictive features for dysbiosis were discovered between donors, but it is reasonable to expect that a larger sample set with repeated examples of similar forms of dysbiosis would be even more informative. Related to this objection is the method used to identify which microbiomes are 'Dysbiosis' and which are 'Healthy'. Without further information, which was unavailable from the David et al. publication, it is difficult to identity the specific criteria by which a microbiome was labeled as Healthy or Dysbiosis, whether some of those microbiomes may be mislabeled, or if there should be additional identifiers, for example 'Pre-dysbiosis'. The lack of additional disruptions to microbiome community structures also limits the opportunities to generate MAP-models from this data. Although in results, a MAP-model was found to be more predictive that an 'Average-abundance' model, the general observation of relatively steady microbiome community structures provided fewer opportunities to model more dynamic microbiomes. The final potential weakness is the lack of opportunity for experimental validation of results. While it would be theoretically possible to ask that 'Donor A' eat an extremely high-fat diet every day for two weeks to validate the dysbiosis model predictions, the likelihood of actually accomplishing this are slim, at best.

7.2.2. Future Work

Many of the solutions to addressing the weaknesses described above would be impossible, impractical, or unethical to implement on human subjects, such as deliberately inducing dysbiosis, finding a very large cohort of genetically similar human volunteers, or dictating a host's diet in such a way that it might be unhealthy or harmful to the volunteer. The best solution to the problems associated with limited human-microbiome datasets is to perform hypothesis-driven experiments using animal models, where many more experimental parameters can be controlled and biological replication may be included. Chapter 6 of this study directly addresses the use of animal-microbiome laboratory models.

7.3. Generating a System-Scale Model of a Mouse Obesity Host-Microbiome Interaction

In this last chapter, a dynamic model of a mouse microbiome was generated, a significant advancement in the prediction of EFP from microbiome community composition was made, and a robust MI model for the prediction of mouse obesity from microbiome community was validated. By combining these models, iMOUSE, a complete mouse obesity HMI system-scale model, was constructed and validated through the successful reproduction of biological experiments *in silico*. Using the system-scale model of mouse HMI, it becomes possible to search a large number of theoretically possible diet parameter combinations to propose a diet that optimizes for a 'Lean' microbiome community structure for any initial microbiome community structure.

The dynamic model of a mouse microbiome community as a function of the host's diet accurately predicts biological observations in a dataset set aside for model validation. Fat, sugar, and fiber diet components were discovered to play a prominent role in shaping the microbiome community structure, as might have been anticipated. More unexpectedly, it was identified that the set of relevant diet parameters for microbiome structure was significantly enriched for vitamin composition. Vitamins, particularly vitamin D (Dimitrov and White 2017, Sirtori et al. 2017), have been implicated in influencing the microbiome community in previous studies. The prominence of dietary vitamin intake in this model suggests that vitamins might be a powerful tool for rationally manipulating microbiome communities. We expect that computational models, including the iMOUSE model constructed in this study, will play a prominent role in determining strategies for manipulating the microbiome through diet and vitamin therapy for patient care.

The TAvP-prediction method, a significant technical advancement of our previously published tool TAP-prediction, calculates an Enzyme Function Profile (EFP) from microbiome community structures. A database of paired shotgun mouse microbiome metagenomes and community structures was combined with a statistical analysis of thousands of sequenced and annotated bacterial genomes to generate highly accurate EFP predictions. The TAvP-prediction tool is expected to become a common and highly valuable application to the scientific community for microbiome analysis.

The prediction approach for host obesity from a microbiome community accurately predicts HMI from the microbiome metabolome. This result is highly significant for a number of reasons. These results once again support our overarching hypothesis that HMI are emergent properties of the microbiome community. A previous meta-analysis of microbiome-obesity studies reports that markers for obesity are not shared across experiments (Walters et al. 2014) and in an analysis of the mouse microbiome dataset we used here, no strong correlations between microbiome community and host obesity were identified (Xiao et al. 2017). In our study, the Obesity Function, calculated from microbiome metabolome, was found to successfully predict obesity across a range of experimental conditions so the limitations in linking microbiome to host phenotype, as identified in multiple microbiome studies and meta-analyses of studies (Fukuda and Ohno 2014, Walters et al. 2014, Xiao et al. 2017), has effectively been overcome.

While each of the modeling approaches developed in Chapter 6 have been proven to be useful analysis tools individually, the true strength of the modeling approaches developed in this aim is their ability to be assembled into a single system-scale model of mouse obesity HMI, the iMOUSE. The iMOUSE model was validated by successfully reproducing *in silico* a series of previous biological experiments: the transplanting of 'Obese' microbiomes into a mouse host predisposed the host for obesity and a gradient of diets from low fat to high fat lead to a similar gradient in propensity for microbiome-associated obesity in mice. Successful validation of the iMOUSE model is a highly significant result. iMOUSE model components (i.e. microbiome community composition prediction, estimation of EFP from community structure, calculation of metagenome from EFP, and prediction of host obesity from microbiome metagenome) were each trained on very different datasets, collected by different laboratories and using mice possessing different gut microbiome communities. For example, the dynamic-MAP model trained on data from one published dataset correctly predicts the behavior of mouse-microbiome interactions in another published dataset, in spite of the fact that the two experiments use mice with very different microbiome community structures. This demonstration that the dynamic MAP-model can accurately extrapolate observation from one experiment to predict another suggests that the computational models have indeed captured key biological phenomenon. The capacity of the complete iMOUSE model to correctly predict an important biological observation that was not present in any of the training data, i.e. that an ‘Obese’ microbiome predisposes a mouse to obesity regardless of the host’s diet conditions, indicates that the model can generate highly relevant biological predictions and provide valuable hypotheses for experimental validation. Using the validated iMOUSE model, diets for mice for the minimization of obesity and based on initial microbiome community structures were determined with high confidence in the biological relevance of the model’s predictions. While both the ‘Lean’-optimized diets and final microbiome community structures were dependent upon initial microbiome community conditions, the resulting microbiomes shared nearly identical obesity-related Enzyme Function Profiles. The predicted host diets and community microbiome structures are within the ranges of previous biological observations, suggesting that the iMOUSE model has likely generated biologically meaningful results.

7.3.1. Potential Weaknesses

In spite of the general success in predicting mouse microbiome obesity HMIs and in the application of the iMOUSE model to *in silico* experiments, a number of potential weaknesses will need to be addressed in future work. For the dynamic modeling of mouse microbiomes, the ‘Gradient’ diet conditions are not the ideal dataset for modeling dynamic populations. Although the ‘Gradient’ data provided a range of diet parameters, a superior set of experimental diets would matrix diet parameters in a more complicated combination of features than the simpler linear combinations of HF and LF used in the

‘Gradient’ data. This would provide a richer collection of varied diets on which the dynamics MAP-models could be trained. There is no immediate solution to this challenge except through conducting additional biological experiments that are designed at the outset to provide the best dynamic ranges of diet parameters for training dynamic MAP-models.

Although the TAvP-prediction tool has been proven to be highly accurate at predicting mouse microbiome EFPs from community structure, it has been optimized here only for mouse gut microbiome communities. If other microbiome community structures are to be analyzed by this approach, new environment-specific TAP-matrixes will need to be generated. The datasets for this proposed work already exist for a range of microbiome environments (Asshauer et al. 2015) and it is our expectation that the TAvP-prediction tool will soon be applied to other microbiome environments, such as soils, marine environments, and hypersaline microbial mats.

While biological validation of model prediction results are outside of the scope of this study, only additional laboratory experiments can truly determine the relative value of the results presented here. Efforts must be made to connect computational models with laboratory science if the promise of manipulating HMI for the benefit of improving human health is to be achieved.

7.3.2. Proposed Future Work

Results of this Aim propose a number of scientifically valuable applications and new research objectives that can be drawn from these analyses.

An additional modification to the dynamic MAP-model approach is suggested by these results. Metabolome had been proven to be a key driver in microbiome populations and HMI repeatedly throughout these analyses. An additional MAP-model parameter that derives from this observation is:

$$taxa_i^t = \sum_{j=1}^{Diet} w_{j,i} diet_j^t + \sum_{k=1}^{Taxa} w_{k,i} taxa_k^t + \sum_{m=1}^{Metabolism} w_{m,i} taxa_m^t \qquad \textbf{Eq. 7.1}$$

Where the terms are the same as in **Eq. 6.1**, with the addition of new terms for the community metabolome (highlighted in red in **Eq. 7.1**). In this variation of dynamic community modeling, the current time point's community structure is also influenced by the previous time point's microbiome community metabolome. While this is an appealing approach, the data required for this is not present in any of the relevant mouse microbiome datasets: a more varied set of initial microbiome community structures, a well-considered matrix of diet parameters, and varied final microbiome community compositions. Implementation of this model must depend upon the identification or generation of appropriate microbiome experiments.

Also, if these modeling results are to have an impact on human health in a clinical setting, the differences between laboratory mouse models and humans must be bridged. While, as previously stated, there is ample reason to believe that for many applications, a mouse model can provide useful insight into human health and disease, there are significant differences between mouse and human diets, their gastrointestinal systems, and their immune systems (Nguyen et al. 2015). One approach to bridging the laboratory model to clinical practice for microbiome-based interventions is to take computational models built on copious mouse data and use those models as a scaffold for the analysis of sparser but more clinically relevant human data. It may be that through minimal and targeted changes to the mouse data-built computational models, the integrated models such as iMOUSE can be rationally modified to better conform to observations in human subjects. In this way, cooperation between the laboratory scientist, clinical researchers, and computational modelers can leverage all of the available data to generate useful predictive and diagnostic tools for investigating the human microbiome.

Finally, it is our goal to take predictive modeling of HMIs beyond obesity. Microbiomes play a role in cancer, autoimmune diseases, diabetes, and a host of other potentially disruptive forms of dysbiosis. The tools and approaches we have developed and validated here will be highly valuable to future investigations into HMIs and potential microbiome-based clinical intervention strategies. We feel that the results of this analysis will prove to be a very important tool for the general scientific community for the modeling and investigations into HMI.

REFERENCES

Aagaard, K., J. Petrosino, W. Keitel, M. Watson, J. Katancik, N. Garcia, S. Patel, M. Cutting, T. Madden, H. Hamilton, E. Harris, D. Gevers, G. Simone, P. McInnes, and J. Versalovic. 2013. The Human Microbiome Project strategy for comprehensive sampling of the human microbiome and why it matters. *FASEB J* **27**:1012-1022.

Ajouz, H., D. Mukherji, and A. Shamseddine. 2014. Secondary bile acids: an underrecognized cause of colon cancer. *World J Surg Oncol* **12**:164.

Alhede, M., T. Bjarnsholt, and M. Givskov. 2014. Pseudomonas aeruginosa biofilms: mechanisms of immune evasion. *Adv Appl Microbiol* **86**:1-40.

Amiot, M. J., C. Riva, and A. Vinet. 2016. Effects of dietary polyphenols on metabolic syndrome features in humans: a systematic review. *Obesity Reviews* **17**:573-586.

Arnold, D. L., A. Pitman, and R. W. Jackson. 2003. Pathogenicity and other genomic islands in plant pathogenic bacteria. *Mol Plant Pathol* **4**:407-420.

Aroniadis, O. C. and L. J. Brandt. 2014. Intestinal microbiota and the efficacy of fecal microbiota transplantation in gastrointestinal disease. *Gastroenterol Hepatol* (N Y) **10**:230-237.

Arora, T., R. L. Loo, J. Anastasovska, G. R. Gibson, K. M. Tuohy, R. K. Sharma, J. R. Swann, E. R. Deaville, M. L. Sleeth, E. L. Thomas, E. Holmes, J. D. Bell, and G. Frost. 2012. Differential effects of two fermentable carbohydrates on central appetite regulation and body composition. *PLoS One* **7**:e43263.

Asshauer, K. P., B. Wemheuer, R. Daniel, and P. Meinicke. 2015. Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. *Bioinformatics* **31**:2882-2884.

Aw, W. and S. Fukuda. 2015. Toward the comprehensive understanding of the gut ecosystem via metabolomics-based integrated omics approach. *Semin Immunopathol* **37**:5-16.

Back, M., A. Avignon, F. Stanke-Labesque, C. Boegner, V. Attalin, E. Leprieur, and A. Sultan. 2014. Leukotriene Production Is Increased in Abdominal Obesity. *PLoS One* **9**.

Bairoch, A. 2000. The ENZYME database in 2000. Nucleic Acids Res **28**:304-305.

Bajaj, J. S., P. B. Hylemon, J. M. Ridlon, D. M. Heuman, K. Daita, M. B. White, P. Monteith, N. A. Noble, M. Sikaroodi, and P. M. Gillevet. 2012. Colonic mucosal microbiome differs from stool microbiome in cirrhosis and hepatic encephalopathy and is linked to cognition and inflammation. A*merican Journal of Physiology-Gastrointestinal and Liver Physiology* **303**:G675-G685.

Barns, S. M., S. L. Takala, and C. R. Kuske. 1999. Wide distribution and diversity of members of the bacterial kingdom Acidobacterium in the environment. *Applied and Environmental Microbiology* ***65****:1731-1737*.

Bashiardes, S., G. Zilberman-Schapira, and E. Elinav. 2016. Use of Metatranscriptomics in Microbiome Research. *Bioinform Biol Insights* **10**:19-25.

Beards, E., K. Tuohy, and G. Gibson. 2010. Bacterial, SCFA and gas profiles of a range of food ingredients following in vitro fermentation by human colonic microbiota. Anaerobe **16**:420-425.

Bervoets, L., K. Van Hoorenbeeck, I. Kortleven, C. Van Noten, N. Hens, C. Vael, H. Goossens, K. N. Desager, and V. Vankerckhoven. 2013. Differences in gut microbiota composition between obese and lean children: a cross-sectional study. *Gut Pathogen*s **5**.

Bhaskaran, K., I. Douglas, H. Forbes, I. dos-Santos-Silva, D. A. Leon, and L. Smeeth. 2014. Body-mass index and risk of 22 specific cancers: a population-based cohort study of 5.24 million UK adults. *Lancet* **384**:755-765.

Blaut, M. 2002. Relationship of prebiotics and food to intestinal microflora. *European Journal of Nutrition* **41 Suppl 1**:I11-16.

Boesjes, M. and G. Brufau. 2014. Metabolic effects of bile acids in the gut in health and disease. *Curr Med Chem* **21**:2822-2829.

Bolker, B., M. Brooks, C. Clark, S. Geange, J. Poulsen, M. Stevens, and J. White. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends Ecol Evol* **3**:171-193.

Borody, T. J., L. J. Brandt, S. Paramsothy, and G. Agrawal. 2013. Fecal microbiota transplantation: a new standard treatment option for Clostridium difficile infection. *Expert Rev Anti Infect Ther* **11**:447-449.

Borre, Y. E., R. D. Moloney, G. Clarke, T. G. Dinan, and J. F. Cryan. 2014. The impact of microbiota on brain and behavior: mechanisms & therapeutic potential. *Software Tools and Algorithms for Biological Systems* **817**:373-403.

Bosi, E., G. Bacci, A. Mengoni, and M. Fondi. 2017. Perspectives and Challenges in Microbial Communities Metabolic Modeling. *Front Genet* **8**:88.

Bou Saab, J., D. Losa, M. Chanson, and R. Ruez. 2014. Connexins in respiratory and gastrointestinal mucosal immunity. *FEBS Lett* **588**:1288-1296.

Bourne, D., Y. Iida, S. Uthicke, and C. Smith-Keune. 2008. Changes in coral-associated microbial communities during a bleaching event. *The ISME Journal* **2**:350-363.

Bouskill, N. J., D. Eveillard, G. O'Mullan, G. A. Jackson, and B. B. Ward. 2011. Seasonal and annual reoccurrence in betaproteobacterial ammonia-oxidizing bacterial population structure. *Environmental Microbiology* **13**:872-886.

Bowers, R. M., C. L. Lauber, C. Wiedinmyer, M. Hamady, A. G. Hallar, R. Fall, R. Knight, and N. Fierer. 2009. Characterization of airborne microbial communities at a high-elevation site and their potential to act as atmospheric ice nuclei. *Applied and Environmental Microbiology* **75**:5121-5130.

Brandt, L. J. and O. C. Aroniadis. 2013. An overview of fecal microbiota transplantation: techniques, indications, and outcomes. *Gastrointest Endosc* **78**:240-249.

Brandt, L. J., T. J. Borody, and J. Campbell. 2011. Endoscopic fecal microbiota transplantation: "first-line" treatment for severe clostridium difficile infection? *J Clin Gastroenterol* **45**:655-657.

Brown, K., D. DeCoffe, E. Molcan, and D. L. Gibson. 2012a. Diet-induced dysbiosis of the intestinal microbiota and the effects on immunity and disease. *Nutrients* **4**:1095-1119.

Brown, S. D., S. M. Utturkar, D. M. Klingeman, C. M. Johnson, S. L. Martin, M. L. Land, T. Y. S. Lu, C. W. Schadt, M. J. Doktycz, and D. A. Pelletier. 2012b. Twenty-One Genome Sequences from Pseudomonas Species and 19 Genome Sequences from Diverse Bacteria Isolated from the Rhizosphere and Endosphere of Populus deltoides. *Journal of Bacteriology* **194**:5991-5993.

Bruggeman, J. and S. A. L. M. Kooijman. 2007. A biodiversity-inspired approach to aquatic ecosystem modeling. *Limnology and Oceanography* **52**:1533-1544.

Bucci, V. and J. B. Xavier. 2014. Towards predictive models of the human gut microbiome. *J Mol Biol* **426**:3907-3916.

Busby, J. R., editor. 1991. *BIOCLIM - A Bioclimatic Analysis and Prediction System*.

Busse, W. W. 1998. Leukotrienes and inflammation. *Am J Respir Crit Care Med* **157**:S210-213.

Calder, P. C. 2013. Feeding the immune system. Proc Nutr Soc **72**:299-309.

Camilleri, M. 2016. High-Fat Diet, Dysbiosis, and Gastrointestinal and Colonic Transit: Is There a Missing Link? *Cell Mol Gastroenterol Hepatol* **2**:257-258.

Cammarota, G., G. Ianiro, R. Cianci, S. Bibbo, A. Gasbarrini, and D. Curro. 2015. The involvement of gut microbiota in inflammatory bowel disease pathogenesis: potential for therapy. *Pharmacol Ther* **149**:191-212.

Cantorna, M. T., K. McDaniel, S. Bora, J. Chen, and J. James. 2014. Vitamin D, immune regulation, the microbiota, and inflammatory bowel disease. *Exp Biol Med* (Maywood) **239**:1524-1530.

Caporaso, J. G., J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Pena, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, and R. Knight. 2010. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7**:335-336.

Carabotti, M., A. Scirocco, M. A. Maselli, and C. Severi. 2015. The gut-brain axis: interactions between enteric microbiota, central and enteric nervous systems. *Ann Gastroenterol* **28**:203-209.

Cardona, C., P. Weisenhorn, C. Henry, and J. A. Gilbert. 2016. Network-based metabolic analysis and microbial community modeling. *Current Opinion in Microbiology* **31**:124-131.

Carmody, R. N., G. K. Gerber, J. M. Luevano, Jr., D. M. Gatti, L. Somes, K. L. Svenson, and P. J. Turnbaugh. 2015. Diet dominates host genotype in shaping the murine gut microbiota. *Cell Host Microbe* **17**:72-84.

Carpenter, G., A. N. Gillison, and J. Winter. 1993. DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. *Biodiversity and Conservation*:667–680.

CDC. Gram-negative Bacteria Infections in Healthcare Settings. 2011. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention.

CDC. Antibiotic Resistance Threats in the United States, 2013. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention.

Chassaing, B., O. Koren, J. K. Goodrich, A. C. Poole, S. Srinivasan, R. E. Ley, and A. T. Gewirtz. 2015. Dietary emulsifiers impact the mouse gut microbiota promoting colitis and metabolic syndrome. *Nature* **519**:92-96.

Che, D. S., Q. Liu, K. Rasheed, and X. P. Tao. 2011. Decision Tree and Ensemble Learning Algorithms with Their Applications in Bioinformatics. *Software Tools and Algorithms for Biological Systems* **696**:191-199.

Chen, I. A., V. M. Markowitz, K. Chu, K. Palaniappan, E. Szeto, M. Pillay, A. Ratner, J. Huang, E. Andersen, M. Huntemann, N. Varghese, M. Hadjithomas, K. Tennessen, T. Nielsen, N. N. Ivanova, and N. C. Kyrpides. 2017. IMG/M: integrated genome and metagenome comparative data analysis system. *Nucleic Acids Res* **45**:D507-D516.

Clarke, P. H. 1982. The metabolic versatility of pseudomonads. *Antonie Van Leeuwenhoek* **48**:105-130.

Clarke, S. F., E. F. Murphy, O. O'Sullivan, A. J. Lucey, M. Humphreys, A. Hogan, P. Hayes, M. O'Reilly, I. B. Jeffery, R. Wood-Martin, D. M. Kerins, E. Quigley, R. P. Ross, P. W. O'Toole, M. G. Molloy, E. Falvey, F. Shanahan, and P. D. Cotter. 2014. Exercise and associated dietary extremes impact on gut microbial diversity. *Gut* **63**:1913-1920.

Clay, K. 2014. Defensive symbiosis: a microbial perspective. Functional Ecology **28**:293-298.

Cockburn, D. W. and N. M. Koropatkin. 2016. Polysaccharide Degradation by the Intestinal Microbiota and Its Influence on Human Health and Disease. *J Mol Biol* **428**:3230-3252.

Collado, M. C., M. Derrien, E. Isolauri, W. M. de Vos, and S. Salminen. 2007. Intestinal integrity and Akkermansia muciniphila, a mucin-degrading member of the intestinal microbiota present in infants, adults, and the elderly. *Applied and Environmental Microbiology* **73**:7767-7770.

Collado, M. C., S. Rautava, E. Isolauri, and S. Salminen. 2015. Gut microbiota: a source of novel tools to reduce the risk of human disease? *Pediatr Res* **77**:182-188.

Collins, S. M. 2014. A role for the gut microbiota in IBS. Nat Rev Gastroenterol Hepatol **11**:497-505.

Conly, J. M. and K. Stein. 1992. The Production of Menaquinones (Vitamin-K2) by Intestinal Bacteria and Their Role in Maintaining Coagulation Homeostasis. *Progress in Food and Nutrition Science* **16**:307-343.

Consortia. 1998. Clinical Guidelines on the Identification, Evaluation, and Treatment of Overweight and Obesity in Adults--The Evidence Report. National Institutes of Health. *Obes Res* **6 Suppl 2**:51S-209S.

Consortium, H. M. P. 2012. Structure, function and diversity of the healthy human microbiome. *Nature* **486**:207-214.

Cook, D., D. Dreyer, D. Bonnet, M. Howell, E. Nony, and K. VandenBosch. 1995. Transient induction of a peroxidase gene in Medicago truncatula precedes infection by Rhizobium meliloti. *Plant Cel*l **7**:43-55.

Corning, P. A. 2012. The re-emergence of emergence, and the causal role of synergy in emergent evolution. *Synthese* **185**:295-317.

Cox, L. M. and M. J. Blaser. 2015. Antibiotics in early life and obesity. *Nat Rev Endocrinol* **11**:182-190.

Cumming, J. R., C. Zawaski, S. Desai, and F. R. Collart. 2015. Phosphorus disequilibrium in the tripartite plant-ectomycorrhiza-plant growth promoting rhizobacterial association. *Journal of Soil Science and Plant Nutrition* **15**:464-485.

David, L. A., A. C. Materna, J. Friedman, M. I. Campos-Baptista, M. C. Blackburn, A. Perrotta, S. E. Erdman, and E. J. Alm. 2014a. Host lifestyle affects human microbiota on daily timescales. *Genome Biol* **15**:R89.

David, L. A., C. F. Maurice, R. N. Carmody, D. B. Gootenberg, J. E. Button, B. E. Wolfe, A. V. Ling, A. S. Devlin, Y. Varma, M. A. Fischbach, S. B. Biddinger, R. J. Dutton, and P. J. Turnbaugh. 2014b. Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**:559-563.

de Bentzmann, S. and P. Plesiat. 2011. The Pseudomonas aeruginosa opportunistic pathogen and human infections. *Environmental Microbiology* **13**:1655-1665.

Degnan, P. H., M. E. Taga, and A. L. Goodman. 2014. Vitamin B-12 as a Modulator of Gut Microbial Ecology. Cell Metabolism **20**:769-778.

Desai, N., D. Antonopoulos, J. A. Gilbert, E. M. Glass, and F. Meyer. 2012. From genomics to metagenomics. *Curr Opin Biotechnol* **23**:72-76.

Devkota, S. and E. B. Chang. 2013. Nutrition, microbiomes, and intestinal inflammation. *Curr Opin Gastroenterol* **29**:603-607.

Dimitrov, V. and J. H. White. 2017. Vitamin D signaling in intestinal innate immunity and homeostasis. *Mol Cell Endocrinol* **453**:68-78.

Ding, T. and P. D. Schloss. 2014. Dynamics and associations of microbial community types across the human body. Nature **509**:357-360.

Dionne, M. S. and D. S. Schneider. 2008. Models of infectious diseases in the fruit fly Drosophila melanogaster. *Dis Model Mech* **1**:43-49.

Distrutti, E., L. Monaldi, P. Ricci, and S. Fiorucci. 2016. Gut microbiota role in irritable bowel syndrome: New therapeutic strategies. *World J Gastroenterol* **22**:2219-2241.

Djerbi, S., M. Lindskog, L. Arvestad, F. Sterky, and T. T. Teeri. 2005. The genome sequence of black cottonwood (Populus trichocarpa) reveals 18 conserved cellulose synthase (CesA) genes. *Planta* **221**:739-746.

Doyle, M. and L. Beuchat. 2001. F*ood microbiology: Fundamentals and Frontiers.* ASM Press, Washington.

Duncan, S. H., G. E. Lobley, G. Holtrop, J. Ince, A. M. Johnstone, P. Louis, and H. J. Flint. 2008. Human colonic microbiota associated with diet, obesity and weight loss. *Int J Obes* (Lond) **32**:1720-1724.

Dupont, H. L. 2014. Review article: evidence for the role of gut microbiota in irritable bowel syndrome and its potential influence on therapeutic targets. *Aliment Pharmacol Ther* **39**:1033-1042.

Dutton, R. J. and P. J. Turnbaugh. 2012. Taking a metagenomic view of human nutrition. Curr Opin Clin *Nutr Metab Care* **15**:448-454.

Economou, V. and P. Gousia. 2015. Agriculture and food animals as a source of antimicrobial-resistant bacteria. *Infection and Drug Resistance* **8**:49-61.

Edwards, R. A., B. Rodriguez-Brito, L. Wegley, M. Haynes, M. Breitbart, D. M. Peterson, M. O. Saar, S. Alexander, E. C. Alexander, Jr., and F. Rohwer. 2006. Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* **7**:57.

Ezenwa, V. O., N. M. Gerardo, D. W. Inouye, M. Medina, and J. B. Xavier. 2012. Microbiology. Animal behavior and the microbiome. *Science* **338**:198-199.

Fernandez, M., M. Porcel, J. de la Torre, M. A. Molina-Henares, A. Daddaoua, M. A. Llamas, A. Roca, V. Carriel, I. Garzon, J. L. Ramos, M. Alaminos, and E. Duque. 2015. Analysis of the pathogenic potential of nosocomial Pseudomonas putida strains. *Frontiers in Microbiology* **6**:871.

Ferrer, J., C. Prats, and D. Lopez. 2008. Individual-based modelling: an essential tool for microbiology. *J Biol Phys* **34**:19-37.

Ferrer, M., A. Ruiz, F. Lanza, S. B. Haange, A. Oberbach, H. Till, R. Bargiela, C. Campoy, M. T. Segura, M. Richter, M. von Bergen, J. Seifert, and A. Suarez. 2013. Microbiota from the distal guts of lean and obese adolescents exhibit partial functional redundancy besides clear differences in community structure. *Environmental Microbiology* **15**:211-226.

Fierer, N., Z. Liu, M. Rodriguez-Hernandez, R. Knight, M. Henn, and M. T. Hernandez. 2008. Short-term temporal variability in airborne bacterial and fungal populations. *Applied and Environmental Microbiology* **74**:200-207.

Finkelstein, E. A., J. G. Trogdon, J. W. Cohen, and W. Dietz. 2009. Annual medical spending attributable to obesity: payer-and service-specific estimates. Health Aff (Millwood) **28**:w822-831.

Follows, M. J., S. Dutkiewicz, S. Grant, and S. W. Chisholm. 2007. Emergent biogeography of microbial communities in a model ocean. *Science* **315**:1843-1846.

Fond, G., W. Boukouaci, G. Chevalier, A. Regnault, G. Eberl, N. Hamdani, F. Dickerson, A. Macgregor, L. Boyer, A. Dargel, J. Oliveira, R. Tamouza, and M. Leboyer. 2015. The "psychomicrobiotic": Targeting microbiota in major psychiatric disorders: A systematic review. *Pathol Biol* (Paris) **63**:35-42.

Foster, J. A. and K. A. McVey Neufeld. 2013. Gut-brain axis: how the microbiome influences anxiety and depression. *Trends in Neurosciences* **36**:305-312.

Franzosa, E. A., X. C. Morgan, N. Segata, L. Waldron, J. Reyes, A. M. Earl, G. Giannoukos, M. R. Boylan, D. Ciulla, D. Gevers, J. Izard, W. S. Garrett, A. T. Chan, and C. Huttenhower. 2014. Relating the metatranscriptome and metagenome of the human gut. *Proceedings of the National Academy of Sciences of the United States of America* **111**:E2329-2338.

Frasca, D., F. Ferracci, A. Diaz, M. Romero, S. Lechner, and B. B. Blomberg. 2016. Obesity decreases B cell responses in young and elderly individuals. *Obesity* **24**:615-625.

Frasca, D., A. M. Landin, S. C. Lechner, J. G. Ryan, R. Schwartz, R. L. Riley, and B. B. Blomberg. 2008. Aging down-regulates the transcription factor E2A, activation-induced cytidine deaminase, and Ig class switch in human B cells. *J Immunol* **180**:5283-5290.

Frey-Klett, P., P. Burlinson, A. Deveau, M. Barret, M. Tarkka, and A. Sarniguet. 2011. Bacterial-Fungal Interactions: Hyphens between Agricultural, Clinical, Environmental, and Food Microbiologists. *Microbiology and Molecular Biology Reviews* **75**:583-+.

Fukuda, S. and H. Ohno. 2014. Gut microbiome and metabolic diseases. *Semin Immunopathol* **36**:103-114.

Fuller, M. 2012. Determination of protein and amino acid digestibility in foods including implications of gut microbial amino acid synthesis. *The British journal of Nutrition* **108 Suppl 2**:S238-246.

Furet, J. P., L. C. Kong, J. Tap, C. Poitou, A. Basdevant, J. L. Bouillot, D. Mariat, G. Corthier, J. Dore, C. Henegar, S. Rizkalla, and K. Clement. 2010. Differential adaptation of human gut microbiota to bariatric surgery-induced weight loss: links with metabolic and low-grade inflammation markers. *Diabetes* **59**:3049-3057.

Furness, J. B., W. A. Kunze, and N. Clerc. 1999. Nutrient tasting and signaling mechanisms in the gut. II. The intestine as a sensory organ: neural, endocrine, and immune responses. *Am J Physiol* **277**:G922-928.

Garbeva, P. and W. de Boer. 2009. Inter-specific Interactions Between Carbon-limited Soil Bacteria Affect Behavior and Gene Expression. *Microbial Ecology* **58**:36-46.

Garenaux, A., M. Caza, and C. M. Dozois. 2011. The Ins and Outs of siderophore mediated iron uptake by extra-intestinal pathogenic Escherichia coli. Veterinary Microbiology **153**:89-98.

Gilbert, J. A. and C. L. Dupont. 2011. Microbial metagenomics: beyond the genome. *Annual Review of Marine Science* **3**:347-371.

Gilbert, J. A. and M. Hughes. 2011. Gene expression profiling: metatranscriptomics. *Methods in Molecular Biology* **733**:195-205.

Gilbert, J. A., B. Laverock, B. Temperton, S. Thomas, M. Muhling, and M. Hughes. 2011. Metagenomics. *Methods in Molecular Biology* **733**:173-183.

Giles, C. D., P. C. Hsu, A. E. Richardson, M. R. H. Hurst, and J. E. Hill. 2014. Plant assimilation of phosphorus from an insoluble organic form is improved by addition of an organic anion producing Pseudomonas sp. *Soil Biology & Biochemistry* **68**:263-269.

Gjesing, A. P. and O. Pedersen. 2012. Omics'-driven discoveries in prevention and treatment of type 2 diabetes. *European Journal of Clinical Investigation* **42**:579-588.

Global, B. M. I. M. C., E. Di Angelantonio, N. Bhupathiraju Sh, D. Wormser, P. Gao, S. Kaptoge, A. Berrington de Gonzalez, B. J. Cairns, R. Huxley, L. Jackson Ch, G. Joshy, S. Lewington, J. E. Manson, N. Murphy, A. V. Patel, J. M. Samet, M. Woodward, W. Zheng, M. Zhou, N. Bansal, A. Barricarte, B. Carter, J. R. Cerhan, G. D. Smith, X. Fang, O. H. Franco, J. Green, J. Halsey, J. S. Hildebrand, K. J. Jung, R. J. Korda, D. F. McLerran, S. C. Moore, L. M. O'Keeffe, E. Paige, A. Ramond, G. K. Reeves, B. Rolland, C. Sacerdote, N. Sattar, E. Sofianopoulou, J. Stevens, M. Thun, H. Ueshima, L. Yang, Y. D. Yun, P. Willeit, E. Banks, V. Beral, Z. Chen, S. M. Gapstur, M. J. Gunter, P. Hartge, S. H. Jee, T. H. Lam, R. Peto, J. D. Potter, W. C. Willett, S. G. Thompson, J. Danesh, and F. B. Hu. 2016. Body-mass index and all-cause mortality: individual-participant-data meta-analysis of 239 prospective studies in four continents. *Lancet* **388**:776-786.

Gomez-Arango, L. F., H. L. Barrett, H. D. McIntyre, L. K. Callaway, M. Morrison, M. D. Nitert, and S. T. Grp. 2016. Connections Between the Gut Microbiome and Metabolic Hormones in Early Pregnancy in Overweight and Obese Women. *Diabetes* **65**:2214-2223.

Gordon, W., W. Semmett, R. Cable, and M. Morris. 1949. Amino Acid Composition of alpha-Casein and beta-Casein. *J Amer Chem Soc* **71**:3293-3300.

Gottel, N. R., H. F. Castro, M. Kerley, Z. M. Yang, D. A. Pelletier, M. Podar, T. Karpinets, E. Uberbacher, G. A. Tuskan, R. Vilgalys, M. J. Doktycz, and C. W. Schadt. 2011. Distinct Microbial Communities within the Endosphere and Rhizosphere of Populus deltoides Roots across Contrasting Soil Types. *Applied and Environmental Microbiology* **77**:5934-5944.

Gradmann, C. 2014. A spirit of scientific rigour: Koch's postulates in twentieth-century medicine. *Microbes Infect* **16**:885-892.

Gras, A., M. Ginovart, X. Portell, and P. C. Baveye. 2010. Individual-Based Modeling of Carbon and Nitrogen Dynamics in Soils: Parameterization and Sensitivity Analysis of Abiotic Components. *Soil Science* **175**:363-374.

Greer, R. L., A. Morgun, and N. Shulzhenko. 2013. Bridging immunity and lipid metabolism by gut microbiota. *J Allergy Clin Immunol* **132**:253-262; quiz 263.

Gritz, E. C. and V. Bhandari. 2015. The human neonatal gut microbiome: a brief review. *Front Pediatr* **3**:17.

Gupta, S., E. Allen-Vercoe, and E. O. Petrof. 2016. Fecal microbiota transplantation: in perspective. *Therap Adv Gastroenterol* **9**:229-239.

Haange, S. B. and N. Jehmlich. 2016. Proteomic interrogation of the gut microbiota: potential clinical impact. *Expert Review of Proteomics* **13**:535-537.

Haas, D. and G. Defago. 2005. Biological control of soil-borne pathogens by fluorescent pseudomonads. *Nat Rev Microbiol* **3**:307-319.

Hacker, J., B. Hochhut, B. Middendorf, G. Schneider, C. Buchrieser, G. Gottschalk, and U. Dobrindt. 2004. Pathogenomics of mobile genetic elements of toxigenic bacteria. *Int J Med Microbiol* **293**:453-461.

Hastie, T., L. Sleeper, and R. Tibshirani. 1992. Flexible covariate effects in the proportional hazards model. *Breast Cancer Research and Treatment* **22**:241-250.

Hastie, T. and R. Tibshirani. 1990. Exploring the nature of covariate effects in the proportional hazards model. *Biometrics* **46**:1005-1016.

Heikkinen, R. K., M. Luoto, R. Virkkala, R. G. Pearson, and J. H. Korber. 2007. Biotic interactions improve prediction of boreal bird distributions at macro-scales. *Global Ecology and Biogeography* **16**:754-763.

Hennessy, A. A., R. P. Ross, G. F. Fitzgerald, N. Caplice, and C. Stanton. 2014. Role of the gut in modulating lipoprotein metabolism. *Curr Cardiol Rep* **16**:515.

Henry, C. S., H. C. Bernstein, P. Weisenhorn, R. C. Taylor, J. Y. Lee, J. Zucker, and H. S. Song. 2016. Microbial Community Metabolic Modeling: A Community Data-Driven Network Reconstruction. *Journal of Cellular Physiology* **231**:2339-2345.

Henry, C. S., M. DeJongh, A. A. Best, P. M. Frybarger, B. Linsay, and R. L. Stevens. 2010. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat Biotechnol* **28**:977-982.

Henry, C. S., R. Overbeek, F. F. Xia, A. A. Best, E. Glass, J. Gilbert, P. Larsen, R. Edwards, T. Disz, F. Meyer, V. Vonstein, M. DeJongh, D. Bartels, N. Desai, M. D'Souza, S. Devoid, K. P. Keegan, R. Olson, A. Wilke, J. Wilkening, and R. L. Stevens. 2011. Connecting genotype to phenotype in the era of high-throughput sequencing. *Biochimica Et Biophysica Acta-General Subjects* **1810**:967-977.

Henson, M. A. and T. J. Hanly. 2014. Dynamic flux balance analysis for synthetic microbial communities. *IET Syst Biol* **8**:214-229.

Hoegh-Guldberg, O. 2010. Dangerous shifts in ocean ecosystem function? *The ISME Journal* **4**:1090-1092.

Holscher, H. D., J. G. Caporaso, S. Hooda, J. M. Brulc, G. C. Fahey, Jr., and K. S. Swanson. 2015. Fiber supplementation influences phylogenetic structure and functional capacity of the human intestinal microbiome: follow-up of a randomized controlled trial. *Am J Clin Nutr* **101**:55-64.

Hong, P. Y., J. A. Croix, E. Greenberg, H. R. Gaskins, and R. I. Mackie. 2011. Pyrosequencing-based analysis of the mucosal microbiota in healthy individuals reveals ubiquitous bacterial groups and micro-heterogeneity. *PLoS One* **6**:e25042.

Hood, R. R., E. A. Laws, R. A. Armstrong, N. R. Bates, C. W. Brown, C. A. Carlson, F. Chai, S. C. Doney, P. G. Falkowski, R. A. Feely, M. A. M. Friedrichs, M. R. Landry, J. K. Moore, D. M. Nelson, T. L. Richardson, B. Salihoglu, M. Schartau, D. A. Toole, and J. D. Wiggert. 2006. Pelagic functional group modeling: Progress, challenges and prospects. *Deep-Sea Research Part II-Topical Studies in Oceanography* **53**:459-512.

Hortin, G., editor. 2012. *Amino acids, peptides, and proteins*. Elsevier, Philadelphia.

Huang, J. C., X. H. Lin, B. Xue, J. M. Luo, L. J. Gao, Y. Wang, S. Y. Ou, and X. C. Peng. 2016. Impact of polyphenols combined with high-fat diet on rats' gut microbiota. *Journal of Functional Foods* **26**:763-771.

Huang, Z. A., X. Chen, Z. Zhu, H. Liu, G. Y. Yan, Z. H. You, and Z. Wen. 2017. PBHMDA: Path-Based Human Microbe-Disease Association Prediction. *Frontiers in Microbiology* **8**:233.

Hugenholtz, P., B. M. Goebel, and N. R. Pace. 1998. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J Bacteriol* **180**:4765-4774.

Huson, D. H., A. F. Auch, J. Qi, and S. C. Schuster. 2007. MEGAN analysis of metagenomic data. *Genome Res* **17**:377-386.

Huttenhower, C. and D. Gevers and R. Knight and S. Abubucker and J. H. Badger and A. T. Chinwalla and H. H. Creasy and A. M. Earl and M. G. FitzGerald and R. S. Fulton and M. G. Giglio and K. Hallsworth-Pepin and E. A. Lobos and R. Madupu and V. Magrini and J. C. Martin and M. Mitreva and D. M. Muzny and E. J. Sodergren and J. Versalovic and A. M. Wollam and K. C. Worley and J. R. Wortman and S. K. Young and Q. D. Zeng and K. M. Aagaard and O. O. Abolude and E. Allen-Vercoe and E. J. Alm and L. Alvarado and G. L. Andersen and S. Anderson and E. Appelbaum and H. M. Arachchi and G. Armitage and C. A. Arze and T. Ayvaz and C. C. Baker and L. Begg and T. Belachew and V. Bhonagiri and M. Bihan and M. J. Blaser and T. Bloom and V. Bonazzi and J. P. Brooks and G. A. Buck and C. J. Buhay and D. A. Busam and J. L. Campbell and S. R. Canon and B. L. Cantarel and P. S. G. Chain and I. M. A. Chen and L. Chen and S. Chhibba and K. Chu and D. M. Ciulla and J. C. Clemente and S. W. Clifton and S. Conlan and J. Crabtree and M. A. Cutting and N. J. Davidovics and C. C. Davis and T. Z. DeSantis and C. Deal and K. D. Delehaunty and F. E. Dewhirst and E. Deych and Y. Ding and D. J. Dooling and S. P. Dugan and W. M. Dunne and A. S. Durkin and R. C. Edgar and R. L. Erlich and C. N. Farmer and R. M. Farrell and K. Faust and M. Feldgarden and V. M. Felix and S. Fisher and A. A. Fodor and L. J. Forney and L. Foster and V. Di Francesco and J. Friedman and D. C. Friedrich and C. C. Fronick and L. L. Fulton and H. Y. Gao and N. Garcia and G. Giannoukos and C. Giblin and M. Y. Giovanni and J. M. Goldberg and J. Goll and A. Gonzalez and A. Griggs and S. Gujja and S. K. Haake and B. J. Haas and H. A. Hamilton and E. L. Harris and T. A. Hepburn and B. Herter and D. E. Hoffmann and M. E. Holder and C. Howarth and K. H. Huang and S. M. Huse and J. Izard and J. K. Jansson and H. Y. Jiang and C. Jordan and V. Joshi and J. A. Katancik and W. A. Keitel and S. T. Kelley and C. Kells and N. B. King and D. Knights and H. D. H. Kong and O. Koren and S. Koren and K. C. Kota and C. L. Kovar and N. C. Kyrpides and P. S. La Rosa and S. L. Lee and K. P. Lemon and N. Lennon and C. M. Lewis and L. Lewis and R. E. Ley and K. Li and K. Liolios and B. Liu and Y. Liu and C. C. Lo and C. A. Lozupone and R. D. Lunsford and T. Madden and A. A. Mahurkar and P. J. Mannon and E. R. Mardis and V. M. Markowitz and K. Mavromatis and J. M. McCorrison and D. McDonald and J. McEwen and A. L. McGuire and P. McInnes and T. Mehta and K. A. Mihindukulasuriya and J. R. Miller and P. J. Minx and I. Newsham and C. Nusbaum and M. O'Laughlin and J. Orvis and I. Pagani and K. Palaniappan and S. M. Patel and M. Pearson and J. Peterson and M. Podar and C. Pohl and K. S. Pollard and M. Pop and M. E. Priest and L. M. Proctor and X. Qin and J. Raes and J. Ravel and J. G. Reid and M. Rho and R. Rhodes and K. P. Riehle and M. C. Rivera and B. Rodriguez-Mueller and Y. H. Rogers and M. C. Ross and C. Russ and R. K. Sanka and P. Sankar and J. F. Sathirapongsasuti and J. A. Schloss and P. D. Schloss and T. M. Schmidt and M. Scholz and L. Schriml and A. M. Schubert and N. Segata and J. A. Segre and W. D. Shannon and R. R. Sharp and T. J. Sharpton and N. Shenoy and N. U. Sheth and G. A. Simone and I. Singh and C. S. Smillie and J. D. Sobel and D. D. Sommer and P. Spicer and G. G. Sutton and S. M. Sykes and D. G. Tabbaa and M. Thiagarajan and C. M. Tomlinson and M. Torralba and T. J. Treangen and R. M. Truty and T. A. Vishnivetskaya and J. Walker and L. Wang and Z. Y. Wang and D. V. Ward and W. Warren and M. A. Watson and C. Wellington and K. A. Wetterstrand and J. R. White and K. Wilczek-Boney and Y. Q. Wu and K. M. Wylie and T. Wylie and C. Yandava and L. Ye and Y. Z. Ye and S. Yooseph and B. P. Youmans and L. Zhang and Y. J. Zhou and Y. M. Zhu and L. Zoloth and J. D. Zucker and B. W. Birren and R. A. Gibbs and S. K. Highlander and B. A. Methe and K. E. Nelson and J. F. Petrosino and G. M. Weinstock and R. K. Wilson and O. White and H. M. P. Consortiu. 2012. Structure, function and diversity of the healthy human microbiome. *Nature* **486**:207-214.

Jackson, R. W., G. M. Preston, and P. B. Rainey. 2005. Genetic characterization of Pseudomonas fluorescens SBW25 rsp gene expression in the phytosphere and in vitro. *J Bacteriol* **187**:8477-8488.

Jansson, J. K., J. D. Neufeld, M. A. Moran, and J. A. Gilbert. 2011. Omics for understanding microbial functional dynamics. *Environmental microbiology*. doi:10.1111/j.1462-2920.2011.02518.x.

Jefferson, K. K. 2004. What drives bacteria to produce a biofilm? *Fems Microbiology Letter*s **236**:163-173.

Jeschke, J. M. and D. L. Strayer. 2008. Usefulness of bioclimatic models for studying climate change and invasive species. *Annals of the New York Academy of Sciences* **1134**:1-24.

Johansson, M. E., J. K. Gustafsson, J. Holmen-Larsson, K. S. Jabbar, L. Xia, H. Xu, F. K. Ghishan, F. A. Carvalho, A. T. Gewirtz, H. Sjovall, and G. C. Hansson. 2014. Bacteria penetrate the normally impenetrable inner colon mucus layer in both murine colitis models and patients with ulcerative colitis. *Gut* **63**:281-291.

Jovel, J., J. Patterson, W. Wang, N. Hotte, S. O'Keefe, T. Mitchel, T. Perry, D. Kao, A. L. Mason, K. L. Madsen, and G. K. Wong. 2016. Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics. *Frontiers in Microbiology* **7**:459.

Kanehisa, M., S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. 2012. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res **40**:D109-114.

Karlsson, F. H., V. Tremaroli, I. Nookaew, G. Bergstrom, C. J. Behre, B. Fagerberg, J. Nielsen, and F. Backhed. 2013. Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* **498**:99-103.

Kasai, C., K. Sugimoto, I. Moritani, J. Tanaka, Y. Oya, H. Inoue, M. Tameda, K. Shiraki, M. Ito, Y. Takei, and K. Takase. 2015. Comparison of the gut microbiota composition between obese and non-obese individuals in a Japanese population, as analyzed by terminal restriction fragment length polymorphism and next-generation sequencing. *BMC Gastroenterology* **15**.

Kasen, S., P. Cohen, H. Chen, and A. Must. 2008. Obesity and psychopathology in women: a three decade prospective study. *Int J Obes* (Lond) **32**:558-566.

Keegan, K. P., E. M. Glass, and F. Meyer. 2016. MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function. Methods in molecular biology **1399**:207-233.

Keung, W., J. R. Ussher, J. S. Jaswal, M. Raubenheimer, V. H. Lam, C. S. Wagg, and G. D. Lopaschuk. 2013. Inhibition of carnitine palmitoyltransferase-1 activity alleviates insulin resistance in diet-induced obese mice. *Diabetes* **62**:711-720.

Khoruts, A. and M. J. Sadowsky. 2016. Understanding the mechanisms of faecal microbiota transplantation. *Nat Rev Gastroenterol Hepatol* **13**:508-516.

Kipanyula, M. J., P. F. Seke Etet, L. Vecchio, M. Farahna, E. N. Nukenine, and A. H. Nwabo Kamdje. 2013. Signaling pathways bridging microbial-triggered inflammation and cancer. *Cell Signa*l **25**:403-416.

Kobyliak, N., C. Conte, G. Cammarota, A. P. Haley, I. Styriak, L. Gaspar, J. Fusek, L. Rodrigo, and P. Kruzliak. 2016. Probiotics in prevention and treatment of obesity: a critical view. *Nutr Metab* (Lond) **13**:14.

Kostic, A. D., M. R. Howitt, and W. S. Garrett. 2013. Exploring host-microbiota interactions in animal models and humans. *Genes Dev* **27**:701-718.

Kostic, A. D., R. J. Xavier, and D. Gevers. 2014. The microbiome in inflammatory bowel disease: current status and the future ahead. *Gastroenterology* **146**:1489-1499.

Krause, L., N. N. Diaz, A. Goesmann, S. Kelley, T. W. Nattkemper, F. Rohwer, R. A. Edwards, and J. Stoye. 2008. Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Res* **36**:2230-2239.

Kurek, J., J. L. Kirk, D. C. Muir, X. Wang, M. S. Evans, and J. P. Smol. 2013. Legacy of a half century of Athabasca oil sands development recorded by lake ecosystems. *Proceedings of the National Academy of Sciences of the United States of America* **110**:1761-1766.

Langille, M. G. I., J. Zaneveld, J. G. Caporaso, D. McDonald, D. Knights, J. A. Reyes, J. C. Clemente, D. E. Burkepile, R. L. V. Thurber, R. Knight, R. G. Beiko, and C. Huttenhower. 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology* **31**:814.

Larque, E., M. Sabater-Molina, and S. Zamora. 2007. Biological significance of dietary polyamines. *Nutrition* **23**:87-95.

Larsen, N., F. K. Vogensen, F. W. van den Berg, D. S. Nielsen, A. S. Andreasen, B. K. Pedersen, W. A. Al-Soud, S. J. Sorensen, L. H. Hansen, and M. Jakobsen. 2010. Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. *PLoS One* **5**:e9085.

Larsen, P., F. Collart, D. Field, F. Meyer, K. Keegan, C. Henry, J. McGrath, J. Quinn, and J. Gilbert. 2011. Predicted Relative Metabolomic Turnover (PRMT): determining metabolic turnover from a coastal marine metagenomic dataset. *Microbial Informatics and Experimentation* **1:4**.

Larsen, P., F. Dawn, and G. JA. 2012a. Predicting bacterial community assemblages using MAP, a novel bioclimatic model. *Nature Biotechnology* **9**:621-625.

Larsen, P., Y. Hamada, and J. Gilbert. 2012b. Modeling microbial communities: Current, developing, and future technologies for predicting microbial community interaction. *Journal of Biotechnolog*y **160**:17-24.

Larsen, P. E., F. R. Collart, and Y. Dai. 2014. Using Metabolomic and Transportomic Modeling and Machine Learning to Identify Putative Novel Therapeutic Targets for Antibiotic Resistant Pseudomonad Infections. 2014 *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (Embc):314-317.

Larsen, P. E., F. R. Collart, and Y. Dai. 2015a. Predicting Ecological Roles in the Rhizosphere Using Metabolome and Transportome Modeling. *PLoS One* **10**.

Larsen, P. E. and Y. Dai. 2015. Metabolome of human gut microbiome is predictive of host dysbiosis. *Gigascience* **4**.

Larsen, P. E., D. Field, and J. A. Gilbert. 2012c. Predicting bacterial community assemblages using an artificial neural network approach. *Nature Methods* **9**:621.

Larsen, P. E., S. M. Gibbons, and J. A. Gilbert. 2012d. Modeling microbial community structure and functional diversity across time and space. *FEMS Microbiology Letters* **332**:91-98.

Larsen, P. E., N. Scott, A. F. Post, D. Field, R. Knight, Y. Hamada, and J. A. Gilbert. 2015b. Satellite remote sensing data can be used to model marine microbial metabolite turnover. *ISME Journal* **9**:166-179.

Larsen, P. E., A. Sreedasyam, G. Trivedi, S. Desai, Y. Dai, L. J. Cseke, and F. R. Collart. 2016. Multi-Omics Approach Identifies Molecular Mechanisms of Plant-Fungus Mycorrhizal Interaction. *Frontiers in Plant Science* **6**.

Lax, S., N. Sangwan, D. Smith, P. Larsen, K. M. Handley, M. Richardson, K. Guyton, M. Krezalek, B. D. Shogan, J. Defazio, I. Flemming, B. Shakhsheer, S. Weber, E. Landon, S. Garcia-Houchins, J. Siegel, J. Alverdy, R. Knight, B. Stephens, and J. A. Gilbert. 2017. Bacterial colonization and succession in a newly opened hospital. *Sci Transl Med* **9**.

Lax, S., D. P. Smith, J. Hampton-Marcell, S. M. Owens, K. M. Handley, N. M. Scott, S. M. Gibbons, P. Larsen, B. D. Shogan, S. Weiss, J. L. Metcalf, L. K. Ursell, Y. Vazquez-Baeza, W. Van Treuren, N. A. Hasan, M. K. Gibson, R. Colwell, G. Dantas, R. Knight, and J. A. Gilbert. 2014. Longitudinal analysis of microbial interaction between humans and the indoor environment. *Science* **345**:1048-1052.

Le Chatelier, E., T. Nielsen, J. Qin, E. Prifti, F. Hildebrand, G. Falony, M. Almeida, M. Arumugam, J. M. Batto, S. Kennedy, P. Leonard, J. Li, K. Burgdorf, N. Grarup, T. Jorgensen, I. Brandslund, H. B. Nielsen, A. S. Juncker, M. Bertalan, F. Levenez, N. Pons, S. Rasmussen, S. Sunagawa, J. Tap, S. Tims, E. G. Zoetendal, S. Brunak, K. Clement, J. Dore, M. Kleerebezem, K. Kristiansen, P. Renault, T. Sicheritz-Ponten, W. M. de Vos, J. D. Zucker, J. Raes, T. Hansen, P. Bork, J. Wang, S. D. Ehrlich, and O. Pedersen. 2013. Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**:541-546.

Lennerz, B. S., S. B. Vafai, N. F. Delaney, C. B. Clish, A. A. Deik, K. A. Pierce, D. S. Ludwig, and V. K. Mootha. 2015. Effects of sodium benzoate, a widely used food preservative, on glucose homeostasis and metabolic profiles in humans. *Molecular Genetics and Metabolism* **114**:73-79.

Lesmes, U., E. J. Beards, G. R. Gibson, K. M. Tuohy, and E. Shimoni. 2008. Effects of resistant starch type III polymorphs on human colon microbiota and short chain fatty acids in human gut models. *J Agric Food Chem* **56**:5415-5421.

Leung, J., B. Burke, D. Ford, G. Garvin, C. Korn, C. Sulis, and N. Bhadelia. 2013. Possible Association between Obesity and Clostridium difficile Infection. Emerging Infectious Diseases **19**:1791-1796.

Li, P., D. Y. Oh, G. Bandyopadhyay, W. S. Lagakos, S. Talukdar, O. Osborn, A. Johnson, H. Chung, M. Maris, J. M. Ofrecio, S. Taguchi, M. Lu, and J. M. Olefsky. 2015. LTB4 promotes insulin resistance in obese mice by acting on macrophages, hepatocytes and myocytes. *Nat Med* **21**:239-247.

Lister, P. D., D. J. Wolter, and N. D. Hanson. 2009. Antibacterial-resistant Pseudomonas aeruginosa: clinical impact and complex regulation of chromosomally encoded resistance mechanisms. *Clin Microbiol Rev* **22**:582-610.

Liu, B., T. Gibbons, M. Ghodsi, T. Treangen, and M. Pop. 2011. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. *BMC Genomics* **12 Suppl 2**:S4.

Liu, R., J. Hong, X. Xu, Q. Feng, D. Zhang, Y. Gu, J. Shi, S. Zhao, W. Liu, X. Wang, H. Xia, Z. Liu, B. Cui, P. Liang, L. Xi, J. Jin, X. Ying, X. Zhao, W. Li, H. Jia, Z. Lan, F. Li, R. Wang, Y. Sun, M. Yang, Y. Shen, Z. Jie, J. Li, X. Chen, H. Zhong, H. Xie, Y. Zhang, W. Gu, X. Deng, B. Shen, H. Yang, G. Xu, Y. Bi, S. Lai, J. Wang, L. Qi, L. Madsen, G. Ning, K. Kristiansen, and W. Wang. 2017. Gut microbiome and serum metabolome alterations in obesity and after weight-loss intervention. *Nat Med* **23**:859-868.

Liu, Z. B., Z. C. Chen, H. W. Guo, D. P. He, H. R. Zhao, Z. Y. Wang, W. Zhang, L. Liao, C. Zhang, and L. Ni. 2016. The modulatory effect of infusions of green tea, oolong tea, and black tea on gut microbiota in high-fat-induced obese mice. *Food & Function* **7**:4869-4879.

Louca, S., A. K. Hawley, S. Katsev, M. Torres-Beltran, M. P. Bhatia, S. Kheirandish, C. C. Michiels, D. Capelle, G. Lavik, M. Doebeli, S. A. Crowe, and S. J. Hallam. 2016. Integrating biogeochemistry with multiomic sequence information in a model oxygen minimum zone. *Proceedings of the National Academy of Sciences of the United States of America* **113**:E5925-E5933.

Lozupone, C. A., M. Hamady, S. T. Kelley, and R. Knight. 2007. Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Applied and Environmental Microbiology* **73**:1576-1585.

Lugtenberg, B. J., L. Dekkers, and G. V. Bloemberg. 2001. Molecular determinants of rhizosphere colonization by Pseudomonas. *Annu Rev Phytopathol* **39**:461-490.

Luna, R. A. and J. A. Foster. 2015. Gut brain axis: diet microbiota interactions and implications for modulation of anxiety and depression. *Curr Opin Biotechnol* **32**:35-41.

Mandal, R. S., S. Saha, and S. Das. 2015. Metagenomic surveys of gut microbiota. *Genomics Proteomics Bioinformatics* **13**:148-158.

Manichanh, C., N. Borruel, F. Casellas, and F. Guarner. 2012. The gut microbiota in IBD. *Nat Rev Gastroenterol Hepatol* **9**:599-608.

Mansfield, J., S. Genin, S. Magori, V. Citovsky, M. Sriariyanum, P. Ronald, M. Dow, V. Verdier, S. V. Beer, M. A. Machado, I. Toth, G. Salmond, and G. D. Foster. 2012. Top 10 plant pathogenic bacteria in molecular plant pathology. Mol Plant Pathol **13**:614-629.

Mao, X. Z., T. Cai, J. G. Olyarchuk, and L. P. Wei. 2005. Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics* **21**:3787-3793.

Marotz, C. A. and A. Zarrinpar. 2016. Treating Obesity and Metabolic Syndrome with Fecal Microbiota Transplantation. *Yale J Biol Med* **89**:383-388.

Martinez-Clemente, M., J. Claria, and E. Titos. 2011. The 5-lipoxygenase/leukotriene pathway in obesity, insulin resistance, and fatty liver disease. Curr Opin Clin Nutr Metab Care **14**:347-353.

Mason, O. U., N. M. Scott, A. Gonzalez, A. Robbins-Pianka, J. Baelum, J. Kimbrel, N. J. Bouskill, E. Prestat, S. Borglin, D. C. Joyner, J. L. Fortney, D. Jurelevicius, W. T. Stringfellow, L. Alvarez-Cohen, T. C. Hazen, R. Knight, J. A. Gilbert, and J. K. Jansson. 2014. Metagenomics reveals sediment microbial community response to Deepwater Horizon oil spill. *The ISME Journal* **8**:1464-1475.

Maynard, C. L., C. O. Elson, R. D. Hatton, and C. T. Weaver. 2012. Reciprocal interactions of the intestinal microbiota and immune system. Nature **489**:231-241.

McFall-Ngai, M. 2007. Adaptive immunity: care for the community. Nature **445**:153.

McKnight, G. M., C. W. Duncan, C. Leifert, and M. H. Golden. 1999. Dietary nitrate in man: friend or foe? British Journal of Nutrition **81**:349-358.

McLean, M. H., D. Dieguez, L. M. Miller, and H. A. Young. 2015. Does the microbiota play a role in the pathogenesis of autoimmune diseases? *Gut* **64**:332-341.

McQuarrie, D. A. 2000. Statistical mechanics. University Science Books, Sausalito, Calif.

Merico, A., J. Bruggeman, and K. Wirtz. 2009. A trait-based approach for downscaling complexity in plankton ecosystem models. *Ecological Modelling* **220**:3001-3010.

Metcalf, J. L., Z. Z. Xu, S. Weiss, S. Lax, W. Van Treuren, E. R. Hyde, S. J. Song, A. Amir, P. Larsen, N. Sangwan, D. Haarmann, G. C. Humphrey, G. Ackermann, L. R. Thompson, C. Lauber, A. Bibat, C. Nicholas, M. J. Gebert, J. F. Petrosino, S. C. Reed, J. A. Gilbert, A. M. Lynne, S. R. Bucheli, D. O. Carter, and R. Knight. 2016. Microbial community assembly and metabolic function during mammalian corpse decomposition. *Science* **351**:158-162.

Metges, C. C. 2000. Contribution of microbial amino acids to amino acid homeostasis of the host. *J Nutr* **130**:1857S-1864S.

Metges, C. C. and K. J. Petzke. 2005. Utilization of essential amino acids synthesized in the intestinal microbiota of monogastric mammals. *The British Journal of Nutrition* **94**:621-622.

Meyer, F., D. Paarmann, M. D'Souza, R. Olson, E. M. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening, and R. A. Edwards. 2008. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**:386.

Million, M., M. Maraninchi, M. Henry, F. Armougom, H. Richet, P. Carrieri, R. Valero, D. Raccah, B. Vialettes, and D. Raoult. 2012. Obesity-associated gut microbiota is enriched in Lactobacillus reuteri and depleted in Bifidobacterium animalis and Methanobrevibacter smithii. I*nt J Obes* (Lond) **36**:817-825.

Mitra, S., P. Rupek, D. C. Richter, T. Urich, J. A. Gilbert, F. Meyer, A. Wilke, and D. H. Huson. 2011. Functional analysis of metagenomes and metatranscriptomes using SEED and KEGG. *BMC Bioinformatics* **12 Suppl 1**:S21.

Moran, C. P. and F. Shanahan. 2014. Gut microbiota and obesity: role in aetiology and potential therapeutic target. *Best Pract Res Clin Gastroenterol* **28**:585-597.

Morris, B. E., R. Henneberger, H. Huber, and C. Moissl-Eichinger. 2013. Microbial syntrophy: interaction for the common good. *FEMS Microbiol Rev* **37**:384-406.

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

Murphy, E. A., K. T. Velazquez, and K. M. Herbert. 2015. Influence of high-fat diet on gut microbiota: a driving force for chronic disease risk. *Curr Opin Clin Nutr Metab Care* **18**:515-520.

Nguyen, T. L. A., S. Vieira-Silva, A. Liston, and J. Raes. 2015. How informative is the mouse for human gut microbiota research? *Disease Models & Mechanisms* **8**:1-16.

Noecker, C., C. P. McNally, A. Eng, and E. Borenstein. 2017. High-resolution characterization of the human microbiome. *Transl Res* **179**:7-23.

Nyholm, S. V. and M. J. McFall-Ngai. 2003. Dominance of Vibrio fischeri in secreted mucus outside the light organ of Euprymna scolopes: the first site of symbiont specificity. *Applied and Environmental Microbiology* **69**:3932-3937.

Nyholm, S. V. and M. J. McFall-Ngai. 2004. The winnowing: establishing the squid-vibrio symbiosis. *Nat Rev Microbiol* **2**:632-642.

Nyholm, S. V., E. V. Stabb, E. G. Ruby, and M. J. McFall-Ngai. 2000. Establishment of an animal-bacterial association: recruiting symbiotic vibrios from the environment. *Proceedings of the National Academy of Sciences of the United States of America* **97**:10231-10235.

O'Callaghan, D. and A. Vergunst. 2010. Non-mammalian animal models to study infectious disease: worms or fly fishing? *Current Opinion in Microbiology* **13**:79-85.

O'Donnell, A. G., I. M. Young, S. P. Rushton, M. D. Shirley, and J. W. Crawford. 2007. Visualization, modelling and prediction in soil microbiology. Nat Rev Microbiol **5**:689-699.

O'dor, R. K., K. Fennel, and E. Vanden Berghe. 2009. A one ocean model of biodiversity. *Deep-Sea Research Part II-Topical Studies in Oceanography* **56**:1816-1823.

Ogata, H., S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa. 1999. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **27**:29-34.

Overballe-Petersen, S., K. Harms, L. A. A. Orlando, J. V. M. Mayar, S. Rasmussen, T. W. Dahl, M. T. Rosing, A. M. Poole, T. Sicheritz-Ponten, S. Brunak, S. Inselmann, J. De Vries, W. Wackernagel, O. G. Pybus, R. Nielsen, P. J. Johnsen, K. M. Nielsen, and E. Willerslev. 2013. Bacterial natural transformation by highly fragmented and damaged DNA. *Proceedings of the National Academy of Sciences of the United States of America* **110**:19860-19865.

Oves-Costales, D., N. Kadi, and G. L. Challis. 2009. The long-overlooked enzymology of a nonribosomal peptide synthetase-independent pathway for virulence-conferring siderophore biosynthesis. *Chem Commun* (Camb):6530-6541.

Palleroni, N. 1992. Introduction to the Pseudomonadaceae. Pages 3071-3085 *in* A. Balows, H. Truper, M. Dworkin, W. Harder, and K. Schleifer, editors. *The Prokaryotes, A Handbook on the Biology of Bacteria, Ecophysiology, Isolation, Identification and Application*s. Springer, New York.

Patil, D. P., D. P. Dhotre, S. G. Chavan, A. Sultan, D. S. Jain, V. B. Lanjekar, J. Gangawani, P. S. Shah, J. S. Todkar, S. Shah, D. R. Ranade, M. S. Patole, and Y. S. Shouche. 2012. Molecular analysis of gut microbiota in obesity among Indian individuals. *Journal of Biosciences* **37**:647-657.

Portune, K. J., M. Beaumont, A. M. Davila, D. Tome, F. Blachier, and Y. Sanz. 2016. Gut microbiota role in dietary protein metabolism and health-related outcomes: The two sides of the coin. *Trends in Food Science & Technology* **57**:213-232.

Putignani, L., F. Del Chierico, A. Petrucca, P. Vernocchi, and B. Dallapiccola. 2014. The human gut microbiota: a dynamic interplay with the host from birth to senescence settled during childhood. *Pediatr Res* **76**:2-10.

Qin, J., R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, D. R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J. M. Batto, T. Hansen, D. Le Paslier, A. Linneberg, H. B. Nielsen, E. Pelletier, P. Renault, T. Sicheritz-Ponten, K. Turner, H. Zhu, C. Yu, M. Jian, Y. Zhou, Y. Li, X. Zhang, N. Qin, H. Yang, J. Wang, S. Brunak, J. Dore, F. Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, P. Bork, and S. D. Ehrlich. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**:59-65.

Qin, J., Y. Li, Z. Cai, S. Li, J. Zhu, F. Zhang, S. Liang, W. Zhang, Y. Guan, D. Shen, Y. Peng, D. Zhang, Z. Jie, W. Wu, Y. Qin, W. Xue, J. Li, L. Han, D. Lu, P. Wu, Y. Dai, X. Sun, Z. Li, A. Tang, S. Zhong, X. Li, W. Chen, R. Xu, M. Wang, Q. Feng, M. Gong, J. Yu, Y. Zhang, M. Zhang, T. Hansen, G. Sanchez, J. Raes, G. Falony, S. Okuda, M. Almeida, E. LeChatelier, P. Renault, N. Pons, J. M. Batto, Z. Zhang, H. Chen, R. Yang, W. Zheng, H. Yang, J. Wang, S. D. Ehrlich, R. Nielsen, O. Pedersen, and K. Kristiansen. 2012. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**:55-60.

Quast, C., E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glockner. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* **41**:D590-D596.

Quiles, F. and F. Humbert. 2014. On the production of glycogen by Pseudomonas fluorescens during biofilm development: an in situ study by attenuated total reflection-infrared with chemometrics. *Biofouling* **30**:709-718.

R-Project. The R Project for Statistical Computing.

Ramakrishna, B. S. 2013. Role of the gut microbiota in human nutrition and metabolism. *J Gastroenterol Hepatol* **28 Suppl 4**:9-17.

Rial, S. A., A. D. Karelis, K. F. Bergeron, and C. Mounier. 2016. Gut Microbiota and Metabolic Health: The Potential Beneficial Effects of a Medium Chain Triglyceride Diet in Obese Individuals. *Nutrients* **8**.

Ridaura, V. K., J. J. Faith, F. E. Rey, J. Cheng, A. E. Duncan, A. L. Kau, N. W. Griffin, V. Lombard, B. Henrissat, J. R. Bain, M. J. Muehlbauer, O. Ilkayeva, C. F. Semenkovich, K. Funai, D. K. Hayashi, B. J. Lyle, M. C. Martini, L. K. Ursell, J. C. Clemente, W. Van Treuren, W. A. Walters, R. Knight, C. B. Newgard, A. C. Heath, and J. I. Gordon. 2013. Gut microbiota from twins discordant for obesity modulate metabolism in mice. *Science* **341**:1241214.

Ridgway, H. F., J. Safarik, D. Phipps, P. Carl, and D. Clark. 1990. Identification and catabolic activity of well-derived gasoline-degrading bacteria from a contaminated aquifer. *Applied and Environmental Microbiology* **56**:3565-3575.

Ridlon, J. M., S. Ikegawa, J. M. P. Alves, B. Zhou, A. Kobayashi, T. Iida, K. Mitamura, G. Tanabe, M. Serrano, A. De Guzman, P. Cooper, G. A. Buck, and P. B. Hylemon. 2013. Clostridium scindens: a human gut microbe with a high potential to convert glucocorticoids into androgens. *Journal of Lipid Research* **54**:2437-2449.

Risto K. Heikkinen, M. L., Miguel B. Araújo, Raimo Virkkala, Wilfried Thuiller and Martin T. Sykes. 2006. Methods and uncertainties in bioclimatic envelope modeling under climate change. *Progress in Physical Geograph*y **30**:751-777.

Rojo, D., A. Hevia, R. Bargiela, P. Lopez, A. Cuervo, S. Gonzalez, A. Suarez, B. Sanchez, M. Martinez-Martinez, C. Milani, M. Ventura, C. Barbas, A. Moya, A. Suarez, A. Margolles, and M. Ferrer. 2015. Ranking the impact of human health disorders on gut metabolism: Systemic lupus erythematosus and obesity as study cases. *Scientific Reports* **5**.

Romano-Keeler, J. and J. H. Weitkamp. 2015. Maternal influences on fetal microbial colonization and immune development. *Pediatr Res* **77**:189-195.

Rooks, M. G. and W. S. Garrett. 2016. Gut microbiota, metabolites and host immunity. *Nat Rev Immunol* **16**:341-352.

Rooks, M. G., P. Veiga, L. H. Wardwell-Scott, T. Tickle, N. Segata, M. Michaud, C. A. Gallini, C. Beal, J. E. van Hylckama-Vlieg, S. A. Ballal, X. C. Morgan, J. N. Glickman, D. Gevers, C. Huttenhower, and W. S. Garrett. 2014. Gut microbiome composition and function in experimental colitis during active disease and treatment-induced remission. *The ISME Journal* **8**:1403-1417.

Rosenbaum, M., R. Knight, and R. L. Leibel. 2015. The gut microbiota in human energy homeostasis and obesity. *Trends Endocrinol Metab* **26**:493-501.

Rossi, M., A. Amaretti, and S. Raimondi. 2011. Folate production by probiotic bacteria. *Nutrients* **3**:118-134.

Round, J. L. and S. K. Mazmanian. 2009. The gut microbiota shapes intestinal immune responses during health and disease. *Nat Rev Immunol* **9**:313-323.

Rubin, T. A., C. E. Gessert, J. Aas, and J. S. Bakken. 2013. Fecal microbiome transplantation for recurrent Clostridium difficile infection: report on a case series. *Anaerobe* **19**:22-26.

Rudrappa, T., M. L. Biedrzycki, and H. P. Bais. 2008. Causes and consequences of plant-associated biofilms. *FEMS Microbiology Ecology* **64**:153-166.

Russell, W. R., S. H. Duncan, L. Scobbie, G. Duncan, L. Cantlay, A. G. Calder, S. E. Anderson, and H. J. Flint. 2013. Major phenylpropanoid-derived metabolites in the human gut can arise from microbial fermentation of protein. Molecular Nutrition & Food Research **57**:523-535.

Ryan, D. and M. Heaner. 2014. Guidelines (2013) for managing overweight and obesity in adults. Preface to the full report. *Obesity* **22 Suppl 2**:S1-3.

Ryan, R. P., K. Germaine, A. Franks, D. J. Ryan, and D. N. Dowling. 2008. Bacterial endophytes: recent developments and applications. *FEMS Microbiology Letters* **278**:1-9.

Sands, D. C., M. N. Schroth, and D. C. Hildebrand. 1970. Taxonomy of phytopathogenic pseudomonads. *J Bacteriol* **101**:9-23.

Sanz, Y. and A. Moya-Perez. 2014. Microbiota, inflammation and obesity. *Software Tools and Algorithms for Biological Systems* **817**:291-317.

Scheffer, M., J. M. Baveco, D. L. Deangelis, K. A. Rose, and E. H. Vannes. 1995. Super-Individuals a Simple Solution for Modeling Large Populations on an Individual Basis. *Ecological Modelling* **80**:161-170.

Schloss, P. D., S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. Van Horn, and C. F. Weber. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology* **75**:7537-7541.

Schulz, M. D., C. Atay, J. Heringer, F. K. Romrig, S. Schwitalla, B. Aydin, P. K. Ziegler, J. Varga, W. Reindl, C. Pommerenke, G. Salinas-Riester, A. Bock, C. Alpert, M. Blaut, S. C. Polson, L. Brandl, T. Kirchner, F. R. Greten, S. W. Polson, and M. C. Arkan. 2014. High-fat-diet-mediated dysbiosis promotes intestinal carcinogenesis independently of obesity. *Nature* **514**:508-512.

Schwabe, R. F. and C. Jobin. 2013. The microbiome and cancer. *Nat Rev Cancer* **13**:800-812.

Schwiertz, A., D. Taras, K. Schafer, S. Beijer, N. A. Bos, C. Donus, and P. D. Hardt. 2010. Microbiota and SCFA in lean and overweight healthy subjects. *Obesity* **18**:190-195.

Sender, R., S. Fuchs, and R. Milo. 2016. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biology* **14**:e1002533.

Shaffer, M., A. J. S. Armstrong, V. V. Phelan, N. Reisdorph, and C. A. Lozupone. 2017. Microbiome and metabolome data integration provides insight into health and disease. *Transl Res*, pii: S1931-5244(17)30232-3.

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

Shen, J., M. S. Obin, and L. Zhao. 2013. The gut microbiota, obesity and insulin resistance. *Mol Aspects Med* **34**:39-58.

Shen, Q., L. Zhao, and K. M. Tuohy. 2012. High-level dietary fibre up-regulates colonic fermentation and relative abundance of saccharolytic bacteria within the human faecal microbiota in vitro. *European Journal of Nutrition* **51**:693-705.

Shi, N., N. Li, X. Duan, and H. Niu. 2017. Interaction between the gut microbiome and mucosal immune system. *Mil Med Res* **4**:14.

Shin, H., Z. Pei, K. A. Martinez, 2nd, J. I. Rivera-Vinas, K. Mendez, H. Cavallin, and M. G. Dominguez-Bello. 2015. The first microbial environment of infants born by C-section: the operating room microbes. *Microbiome* **3**:59.

Shinde, S., J. R. Cumming, F. R. Collart, P. H. Noirot, and P. E. Larsen. 2017. Pseudomonas fluorescens Transportome Is Linked to Strain-Specific Plant Growth Promotion in Aspen Seedlings under Nutrient Stress. *Frontiers in Plant Science* **8**.

Shoaie, S., P. Ghaffari, P. Kovatcheva-Datchary, A. Mardinoglu, P. Sen, E. Pujos-Guillot, T. de Wouters, C. Juste, S. Rizkalla, J. Chilloux, L. Hoyles, J. K. Nicholson, J. Dore, M. E. Dumas, K. Clement, F. Backhed, and J. Nielsen. 2015. Quantifying Diet-Induced Metabolic Changes of the Human Gut Microbiome. *Cell Metabolism* **22**:320-331.

Shtarkman, Y. M., Z. A. Kocer, R. Edgar, R. S. Veerapaneni, T. D'Elia, P. F. Morris, and S. O. Rogers. 2013. Subglacial Lake Vostok (Antarctica) accretion ice contains a diverse set of sequences from aquatic, marine and sediment-inhabiting bacteria and eukarya. *PLoS One* **8**:e67221.

Silby, M. W., A. M. Cerdeno-Tarraga, G. S. Vernikos, S. R. Giddens, R. W. Jackson, G. M. Preston, X. X. Zhang, C. D. Moon, S. M. Gehrig, S. A. Godfrey, C. G. Knight, J. G. Malone, Z. Robinson, A. J. Spiers, S. Harris, G. L. Challis, A. M. Yaxley, D. Harris, K. Seeger, L. Murphy, S. Rutter, R. Squares, M. A. Quail, E. Saunders, K. Mavromatis, T. S. Brettin, S. D. Bentley, J. Hothersall, E. Stephens, C. M. Thomas, J. Parkhill, S. B. Levy, P. B. Rainey, and N. R. Thomson. 2009. Genomic and genetic analyses of diversity and plant interactions of Pseudomonas fluorescens. Genome Biol **10**:R51.

Silby, M. W., C. Winstanley, S. A. Godfrey, S. B. Levy, and R. W. Jackson. 2011. Pseudomonas genomes: diverse and adaptable. *FEMS Microbiol Rev* **35**:652-680.

Sirtori, C. R., C. Pavanello, L. Calabresi, and M. Ruscica. 2017. Nutraceutical Approaches to the Metabolic Syndrome. Ann Med:1-40.

Smirnov, K. S., T. V. Maier, A. Walker, S. S. Heinzmann, S. Forcisi, I. Martinez, J. Walter, and P. Schmitt-Kopplin. 2016. Challenges of metabolomics in human gut microbiota research. *Int J Med Microbiol* **306**:266-279.

Smith, V. A., J. Yu, T. V. Smulders, A. J. Hartemink, and E. D. Jarvis. 2006. Computational inference of neural information flow networks. *PLoS Computational Biology* **2**:e161.

Sonnenburg, J. L. and F. Backhed. 2016. Diet-microbiota interactions as moderators of human metabolism. *Nature* **535**:56-64.

Spor, A., O. Koren, and R. Ley. 2011. Unravelling the effects of the environment and host genotype on the gut microbiome. *Nat Rev Microbiol* **9**:279-290.

Stein, R. R., V. Bucci, N. C. Toussaint, C. G. Buffie, G. Ratsch, E. G. Pamer, C. Sander, and J. B. Xavier. 2013. Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. *PLoS Computational Biology* **9**:e1003388.

Stockwell, D. R. B. and I. R. Noble. 1992. Induction of sets of rules from animal distribution data: A robust and informative method of analysis. Mathematics and Computers in Simulation:385-390.

Stockwell, D. R. B. and D. P. Peters. 1999. The GARP modelling system: Problems and solutions to automated spatial prediction. *International Journal of Geographic Information Systems*:143-158.

Stulberg, E., D. Fravel, L. M. Proctor, D. M. Murray4, J. LoTempio, L. Chrisey, J. Garland, K. Goodwin, J. Graber, M. C. Harris, S. Jackson, M. Mishkind, D. M. Porterfield, and A. Records. 2016. An assessment of US microbiome research. *Nature Microbiology* **1**.

Sung, M. M., T. T. Kim, E. Denou, C. L. M. Soltys, S. M. Hamza, N. J. Byrne, G. Masson, H. Park, D. S. Wishart, K. L. Madsen, J. D. Schertzer, and J. R. B. Dyck. 2017. Improved Glucose Homeostasis in Obese Mice Treated With Resveratrol Is Associated With Alterations in the Gut Microbiome. *Diabetes* **66**:418-425.

Suzuki, S., T. Horinouchi, and C. Furusawa. 2014. Prediction of antibiotic resistance by gene expression profiles. *Nat Commun* **5**:5792.

Takai, K., D. P. Moser, M. DeFlaun, T. C. Onstott, and J. K. Fredrickson. 2001. Archaeal diversity in waters from deep South African gold mines. *Applied and Environmental Microbiology* **67**:5750-5760.

Tamboli, C. P., C. Neut, P. Desreumaux, and J. F. Colombel. 2004. Dysbiosis in inflammatory bowel disease. *Gut* **53**:1-4.

Teske, A. and K. B. Sorensen. 2008. Uncultured archaea in deep marine subsurface sediments: have we caught them all? *The ISME Journal* **2**:3-18.

Thaiss, C. A., S. I. Tav, D. Rothschild, M. T. M. Eijer, M. Levy, C. Moresi, L. Dohnalova, S. Braverman, S. Rozin, S. Malitsky, M. Dori-Bachash, Y. Kuperman, I. Biton, A. Gertler, A. Harmelin, H. Shapiro, Z. Halpern, A. Aharoni, E. Segal, and E. Elinav. 2016. Persistent microbiome alterations modulate the rate of post-dieting weight regain. *Nature* **540**:544.

Theriot, C. M., M. J. Koenigsknecht, P. E. Carlson, Jr., G. E. Hatton, A. M. Nelson, B. Li, G. B. Huffnagle, Z. L. J, and V. B. Young. 2014. Antibiotic-induced shifts in the mouse gut microbiome and metabolome increase susceptibility to Clostridium difficile infection. *Nat Commun* **5**:3114.

Thomas, T., J. Gilbert, and F. Meyer. 2012. Metagenomics **-** a guide from sampling to data analysis. *Microb Inform Exp* **2**:3.

Timm, C. M., A. G. Campbell, S. M. Utturkar, S. R. Jun, R. E. Parales, W. A. Tan, M. S. Robeson, T. Y. Lu, S. Jawdy, S. D. Brown, D. W. Ussery, C. W. Schadt, G. A. Tuskan, M. J. Doktycz, D. J. Weston, and D. A. Pelletier. 2015. Metabolic functions of Pseudomonas fluorescens strains from Populus deltoides depend on rhizosphere or endosphere isolation compartment. *Frontiers in Microbiology* **6**:1118.

Tims, S., C. Derom, D. M. Jonkers, R. Vlietinck, W. H. Saris, M. Kleerebezem, W. M. de Vos, and E. G. Zoetendal. 2013. Microbiota conservation and BMI signatures in adult monozygotic twins. The *ISME Journal* **7**:707-717.

Turnbaugh, P. J., F. Baeckhed, L. Fulton, and J. I. Gordon. 2008. Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome. *Cell Host & Microbe* **3**:213-223.

Turnbaugh, P. J. and J. I. Gordon. 2008. An invitation to the marriage of metagenomics and metabolomics. *Cell* **134**:708-713.

Turnbaugh, P. J., M. Hamady, T. Yatsunenko, B. L. Cantarel, A. Duncan, R. E. Ley, M. L. Sogin, W. J. Jones, B. A. Roe, J. P. Affourtit, M. Egholm, B. Henrissat, A. C. Heath, R. Knight, and J. I. Gordon. 2009a. A core gut microbiome in obese and lean twins. *Nature* **457**:480-484.

Turnbaugh, P. J., R. E. Ley, M. A. Mahowald, V. Magrini, E. R. Mardis, and J. I. Gordon. 2006. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444**:1027-1031.

Turnbaugh, P. J., V. K. Ridaura, J. J. Faith, F. E. Rey, R. Knight, and J. I. Gordon. 2009b. The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Sci Transl Med* **1**:6ra14.

Tuskan, G. A. and S. Difazio and S. Jansson and J. Bohlmann and I. Grigoriev and U. Hellsten and N. Putnam and S. Ralph and S. Rombauts and A. Salamov and J. Schein and L. Sterck and A. Aerts and R. R. Bhalerao and R. P. Bhalerao and D. Blaudez and W. Boerjan and A. Brun and A. Brunner and V. Busov and M. Campbell and J. Carlson and M. Chalot and J. Chapman and G. L. Chen and D. Cooper and P. M. Coutinho and J. Couturier and S. Covert and Q. Cronk and R. Cunningham and J. Davis and S. Degroeve and A. Dejardin and C. Depamphilis and J. Detter and B. Dirks and I. Dubchak and S. Duplessis and J. Ehlting and B. Ellis and K. Gendler and D. Goodstein and M. Gribskov and J. Grimwood and A. Groover and L. Gunter and B. Hamberger and B. Heinze and Y. Helariutta and B. Henrissat and D. Holligan and R. Holt and W. Huang and N. Islam-Faridi and S. Jones and M. Jones-Rhoades and R. Jorgensen and C. Joshi and J. Kangasjarvi and J. Karlsson and C. Kelleher and R. Kirkpatrick and M. Kirst and A. Kohler and U. Kalluri and F. Larimer and J. Leebens-Mack and J. C. Leple and P. Locascio and Y. Lou and S. Lucas and F. Martin and B. Montanini and C. Napoli and D. R. Nelson and C. Nelson and K. Nieminen and O. Nilsson and V. Pereda and G. Peter and R. Philippe and G. Pilate and A. Poliakov and J. Razumovskaya and P. Richardson and C. Rinaldi and K. Ritland and P. Rouze and D. Ryaboy and J. Schmutz and J. Schrader and B. Segerman and H. Shin and A. Siddiqui and F. Sterky and A. Terry and C. J. Tsai and E. Uberbacher and P. Unneberg and J. Vahala and K. Wall and S. Wessler and G. Yang and T. Yin and C. Douglas and M. Marra and G. Sandberg and Y. Van de Peer and D. Rokhsar. 2006. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science* **313**:1596-1604.

Ursell, L. K., H. J. Haiser, W. Van Treuren, N. Garg, L. Reddivari, J. Vanamala, P. C. Dorrestein, P. J. Turnbaugh, and R. Knight. 2014. The intestinal metabolome: an intersection between microbiota and host. *Gastroenterology* **146**:1470-1476.

van Baarlen, P., M. Kleerebezem, and J. M. Wells. 2013. Omics approaches to study host-microbiota interactions. *Current Opinion in Microbiolog*y **16**:270-277.

Varankovich, N. V., M. T. Nickerson, and D. R. Korber. 2015. Probiotic-based strategies for therapeutic and prophylactic use against multiple gastrointestinal diseases. Frontiers in Microbiology **6**.

Varma, A. and B. O. Palsson. 1994a. *Metabolic Flux Balancing - Basic Concepts, Scientific and Practical Use. Bio-Technology* **12**:994-998.

Varma, A. and B. O. Palsson. 1994b. Stoichiometric Flux Balance Models Quantitatively Predict Growth and Metabolic by-Product Secretion in Wild-Type Escherichia-Coli W3110. *Applied and Environmental Microbiology* **60**:3724-3731.

Verberkmoes, N. C., A. L. Russell, M. Shah, A. Godzik, M. Rosenquist, J. Halfvarson, M. G. Lefsrud, J. Apajalahti, C. Tysk, R. L. Hettich, and J. K. Jansson. 2009. Shotgun metaproteomics of the human distal gut microbiota. T*he ISME Journal* **3**:179-189.

Verdam, F. J., S. Fuentes, C. de Jonge, E. G. Zoetendal, R. Erbil, J. W. Greve, W. A. Buurman, W. M. de Vos, and S. S. Rensen. 2013. Human intestinal microbiota composition is associated with local and systemic inflammation in obesity. *Obesity* **21**:E607-E615.

Viaud, S., R. Daillere, I. G. Boneca, P. Lepage, M. J. Pittet, F. Ghiringhelli, G. Trinchieri, R. Goldszmid, and L. Zitvogel. 2014. Harnessing the intestinal microbiome for optimal therapeutic immunomodulation. *Cancer Research* **74**:4217-4221.

Walker, P. A. and K. D. Cocks. 1991. HABITAT: a procedure for modelling a disjoint environmental envelope for a plant or animal species. Global Ecology and Biogeography Letters **1**:108–118.

Walsh, C. J., C. M. Guinane, P. W. O'Toole, and P. D. Cotter. 2014. Beneficial modulation of the gut microbiota. *FEBS Lett* **588**:4120-4130.

Walters, W. A., Z. Xu, and R. Knight. 2014. Meta-analyses of human gut microbes associated with obesity and IBD. *FEBS Letters* **588**:4223-4233.

Wang, W. L., S. Y. Xu, Z. G. Ren, L. Tao, J. W. Jiang, and S. S. Zheng. 2015. Application of metagenomics in the human gut microbiome. World J Gastroenterol **21**:803-814.

Wang, Y. C., K. McPherson, T. Marsh, S. L. Gortmaker, and M. Brown. 2011. Health and economic burden of the projected obesity trends in the USA and the UK. *Lancet* **378**:815-825.

Ward, B. A., M. A. M. Friedrichs, T. R. Anderson, and A. Oschlies. 2010. Parameter optimisation techniques and the problem of underdetermination in marine biogeochemical models. *Journal of Marine Systems* **81**:34-43.

Ward, T. G. and P. C. Trexler. 1958. Gnotobiotics: a new discipline in biological and medical research. *Perspect Biol Med* **1**:447-456.

Whitman, W. B., D. C. Coleman, and W. J. Wiebe. 1998. Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences of the United States of America* **95**:6578-6583.

Wong, H. L., A. Ahmed-Cox, and B. P. Burns. 2016. Molecular Ecology of Hypersaline Microbial Mats: Current Insights and New Directions. *Microorganisms* **4**.

Wood, D. E. and S. L. Salzberg. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* **15**.

Wu, G. D., J. Chen, C. Hoffmann, K. Bittinger, Y. Y. Chen, S. A. Keilbaugh, M. Bewtra, D. Knights, W. A. Walters, R. Knight, R. Sinha, E. Gilroy, K. Gupta, R. Baldassano, L. Nessel, H. Li, F. D. Bushman, and J. D. Lewis. 2011. Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334**:105-108.

Wylie, K. M., R. M. Truty, T. J. Sharpton, K. A. Mihindukulasuriya, Y. Zhou, H. Gao, E. Sodergren, G. M. Weinstock, and K. S. Pollard. 2012. Novel bacterial taxa in the human microbiome. *PLoS One* **7**:e35294.

Xiao, L., Q. Feng, S. Liang, S. B. Sonne, Z. Xia, X. Qiu, X. Li, H. Long, J. Zhang, D. Zhang, C. Liu, Z. Fang, J. Chou, J. Glanville, Q. Hao, D. Kotowska, C. Colding, T. R. Licht, D. Wu, J. Yu, J. J. Sung, Q. Liang, J. Li, H. Jia, Z. Lan, V. Tremaroli, P. Dworzynski, H. B. Nielsen, F. Backhed, J. Dore, E. Le Chatelier, S. D. Ehrlich, J. C. Lin, M. Arumugam, J. Wang, L. Madsen, and K. Kristiansen. 2015. A catalog of the mouse gut metagenome. *Nat Biotechnol* **33**:1103-1108.

Xiao, L., S. B. Sonne, Q. Feng, N. Chen, Z. Xia, X. Li, Z. Fang, D. Zhang, E. Fjaere, L. K. Midtbo, M. Derrien, F. Hugenholtz, L. Tang, J. Li, J. Zhang, C. Liu, Q. Hao, U. B. Vogel, A. Mortensen, M. Kleerebezem, T. R. Licht, H. Yang, J. Wang, Y. Li, M. Arumugam, L. Madsen, and K. Kristiansen. 2017. High-fat feeding rather than obesity drives taxonomical and functional changes in the gut microbiota in mice. *Microbiome* **5**:43.

Xiong, W., P. E. Abraham, Z. Li, C. Pan, and R. L. Hettich. 2015. Microbial metaproteomics for characterizing the range of metabolic functions and activities of human gut microbiota. *Proteomics* **15**:3424-3438.

Yang, H., X. C. Huang, S. M. Fang, W. S. Xin, L. S. Huang, and C. Y. Chen. 2016a. Uncovering the composition of microbial community structure and metagenomics among three gut locations in pigs with distinct fatness. *Scientific Reports* **6**.

Yang, I., E. J. Corwin, P. A. Brennan, S. Jordan, J. R. Murphy, and A. Dunlop. 2016b. The Infant Microbiome: Implications for Infant Health and Neurocognitive Development. *Nurs Res* **65**:76-88.

Yang, Y. and C. Jobin. 2014. Microbial imbalance and intestinal pathologies: connections and contributions. *Dis Model Mech* **7**:1131-1142.

Yi, P. and L. J. Li. 2012. The germfree murine animal: An important animal model for research on the relationship between gut microbiota and the host. *Veterinary Microbiology* **157**:1-7.

Yoo, J. Y. and S. S. Kim. 2016. Probiotics and Prebiotics: Present Status and Future Perspectives on Metabolic Disorders. *Nutrients* **8**:173.

Yoon, S. S., E. K. Kim, and W. J. Lee. 2015. Functional genomic and metagenomic approaches to understanding gut microbiota-animal mutualism. *Current Opinion in Microbiology* **24**:38-46.

Ze, X. L., S. H. Duncan, P. Louis, and H. J. Flint. 2012. Ruminococcus bromii is a keystone species for the degradation of resistant starch in the human colon. *ISME Journal* **6**:1535-1543.

Zhang, C. H., M. H. Zhang, S. Y. Wang, R. J. Han, Y. F. Cao, W. Y. Hua, Y. J. Mao, X. J. Zhang, X. Y. Pang, C. C. Wei, G. P. Zhao, Y. Chen, and L. P. Zhao. 2010. Interactions between gut microbiota, host genetics and diet relevant to development of metabolic syndromes in mice. *ISME Journal* **4**:312-313.

Zhang, K., H. Choi, D. D. Dionysiou, G. A. Sorial, and D. B. Oerther. 2006. Identifying pioneer bacterial species responsible for biofouling membrane bioreactors. *Environmental Microbiology* **8**:433-440.

Zhang, M. and X. J. Yang. 2016. Effects of a high fat diet on intestinal microbiota and gastrointestinal diseases. *World J Gastroenterol* **22**:8905-8909.

APPENDIXES

Appendix A. Pseudomonas Genomes

Pseudomonas Species	Taxonomy Browser Link	Link to NCBI Sequence	Additional references
Pseudomonas aeruginosa B136-33	ID=1280938	RefSeq	
Pseudomonas aeruginosa DK2	ID=1093787	RefSeq	PMID:22672046
Pseudomonas aeruginosa LES431	ID=1408272	RefSeq	PMID:22672046
Pseudomonas aeruginosa M18	ID=941193	RefSeq	PMID:21884571
Pseudomonas aeruginosa NCGM2.S1	ID=1089456	RefSeq	PMID:22123763
Pseudomonas aeruginosa PA7	ID=381754	RefSeq	PMID:20107499
Pseudomonas aeruginosa PAO1	ID=1147787	RefSeq	PMID:10984043
Pseudomonas aeruginosa RP73	ID=1340851	RefSeq	PMID:23908295
Pseudomonas aeruginosa UCBPP-PA14	ID=208963	RefSeq	PMID:17038190
Pseudomonas brassicacearum NFM421	ID=930166	RefSeq	PMID:21515771
Pseudomonas denitrificans ATCC 13867	ID=1294143	RefSeq	PMID:23723394 PMC:2618026/
Pseudomonas entomophila L48	ID=384676	RefSeq	PMID:16699499
Pseudomonas fluorescens A506	ID=1037911	RefSeq	PMID:22792073
Pseudomonas fluorescens F113	ID=1114970	RefSeq	PMID:22328765
Pseudomonas fluorescens Pf0-1	ID=205922	RefSeq	PMID:19432983
Pseudomonas fluorescens SBW25	ID=216595	RefSeq	PMID:19432983
Pseudomonas fulva 12-X	ID=743720	RefSeq	PMID:17668039
Pseudomonas mendocina NK-01	ID=1001585	RefSeq	PMID:21551299
Pseudomonas mendocina ymp	ID=399739	RefSeq	PMID:17384310
Pseudomonas putida ND6	ID=231023	RefSeq	PMID:23046581
Pseudomonas poae RE*1-1-14	ID=1282356	RefSeq	PMID:23516179
Pseudomonas protegens CHA0	ID=1124983	RefSeq	PMID:24762936 SAM:PprotCHA0
Pseudomonas protegens Pf-5	ID=220664	RefSeq	PMID:15980861
Pseudomonas putida BIRD-1	ID=931281	RefSeq	PMID:21183676
Pseudomonas putida DOT-T1E	ID=1196325	RefSeq	PMID:22819823
Pseudomonas putida F1	ID=351746	RefSeq	PMC:3056050/
Pseudomonas putida GB-1	ID=76869	RefSeq	PMC:3056050/
Pseudomonas putida H8234	ID=1331671	RefSeq	PMID:23868128
Pseudomonas putida HB3267	ID=1215088	RefSeq	PMID:24465371
Pseudomonas putida KT2440	ID=160488	RefSeq	PMC:3056050/ PMID:12534463

Pseudomonas putida NBRC 14164	ID=1211579	RefSeq	PMID:24526630
Pseudomonas putida S16	ID=1042876	RefSeq	PMID:21914868
Pseudomonas putida W619	ID=390235	RefSeq	PMC:3056050/
Pseudomonas sp. UW4	ID=1207075	RefSeq	PMC:3596284/
Pseudomonas stutzeri A1501	ID=379731	RefSeq	PMID:18495935
Pseudomonas stutzeri ATCC 17588	ID=96563	RefSeq	PMID:21994926
Pseudomonas stutzeri CCUG 29243	ID=1196835	RefSeq	PMID:23144395
Pseudomonas stutzeri DSM 10701	ID=1123519	RefSeq	PMID:22965097
Pseudomonas stutzeri DSM 4166	ID=996285	RefSeq	PMID:21515765
Pseudomonas stutzeri RCH2	ID=644801	RefSeq	JGI=PsestuRCH2
Pseudomonas syringae pv. phaseolicola 1448A	ID=264730	RefSeq	PMID:16159782
Pseudomonas syringae pv. syringae B728a	ID=205918	RefSeq	PMID:16043691
Pseudomonas syringae pv. tomato DC3000	ID=223283	RefSeq	PMID:12928499

Appendix B. BANJO Parameters for Human Donor A CIN

```
###-------------------------------------------------
### Project information
###-------------------------------------------------

project =      Microbiome Data, Fa, Order + Nutrition
user =                                        PLarsen
dataset =                    30-vars-144-observations
notes =                         Dynamic network prediction

###-------------------------------------------------
### Search component specifications
###-------------------------------------------------

searcherChoice =                           SimAnneal
proposerChoice =                     RandomLocalMove
evaluatorChoice =                            default
deciderChoice =                              default

###-------------------------------------------------
### Input and output locations
###-------------------------------------------------

inputDirectory =              /disk1/Microbiome_BNI/data
observationsFile =                   MicrobiomeFa_Data.txt
outputDirectory =           /disk1/Microbiome_BNI/output
reportFile =                 MicrobiomeOrderFa-q4.@TS@.txt

variablesAreInRows=                     yes
###-------------------------------------------------
### We require this only to validate the input
###-------------------------------------------------

variableCount =                                 30

###-------------------------------------------------
### Pre-processing options
###-------------------------------------------------

discretizationPolicy =                          i4
discretizationExceptions =
createDiscretizationReport =                      no

###-------------------------------------------------
### Network structure properties
###-------------------------------------------------

minMarkovLag =                                     0
maxMarkovLag =                                     1
dbnMandatoryIdentityLags =
equivalentSampleSize =                           1.0
maxParentCount =                                   5
defaultMaxParentCount =                            7

###-------------------------------------------------
```

```
### Network structure properties, optional
###-------------------------------------------------

initialStructureFile =
mustBePresentEdgesFile =
mustNotBePresentEdgesFile =       MicrobiomeFa_NoEdges.txt

###-------------------------------------------------
### Stopping criteria
###-------------------------------------------------

maxTime =                                       12 h
maxProposedNetworks =
maxRestarts =                                  10000
minNetworksBeforeChecking =                     1000

###-------------------------------------------------
### Search monitoring properties
###-------------------------------------------------

nBestNetworks =                                    3
bestNetworksAre =
screenReportingInterval =                       10 m
fileReportingInterval =                         60 m

###-------------------------------------------------
### Parameters used by specific search methods
###-------------------------------------------------

### For simulated annealing:
initialTemperature =                           10000
coolingFactor =                                  0.7
reannealingTemperature =                         800
maxAcceptedNetworksBeforeCooling =              2500
maxProposedNetworksBeforeCooling =             10000
minAcceptedNetworksBeforeReannealing =           500

### For greedy:
minProposedNetworksAfterHighScore =             1000
minProposedNetworksBeforeRestart =              3000
maxProposedNetworksBeforeRestart =              5000
restartWithRandomNetwork =                       yes
maxParentCountForRestart =                         3

###-------------------------------------------------
### Command line user interface options
###-------------------------------------------------

askToVerifySettings =                             no

###-------------------------------------------------
### Post-processing options
###-------------------------------------------------

createDotOutput =                                no
computeInfluenceScores =                         yes
computeConsensusGraph =                          yes
createConsensusGraphAsHtml =                     no
fileNameForTopGraph =                 top.graph.@TS@
fileNameForConsensusGraph =     consensus.graph.@TS@
dotGraphicsFormat =                              jpg
dotFileExtension =                               txt
htmlFileExtension =                             html
```

```
fullPathToDotExecutable = ### C:/Program Files/ATT/Graphviz/bin/dot.exe
timeStampFormat =                   yyyy.MM.dd.HH.mm

###-------------------------------------------------
### Memory management and performance options
###-------------------------------------------------

precomputeLogGamma =                             yes
useCache =                                fastLevel2
cycleCheckingMethod =                            dfs

###-------------------------------------------------
### Misc. options
###-------------------------------------------------

displayMemoryInfo =                              yes
displayStructures =                              yes
```

Appendix C. Human Microbiome CIN Network

rhi	reg	act
rho	reg	act
FIBER	reg	bif
bif	reg	bif
FIBER	reg	cor
SATFAT	reg	cor
bif	reg	cor
cor	reg	cor
cau	reg	cor
FIBER	reg	bac
bif	reg	bac
ery	reg	bac
FAT	reg	lac
FIBER	reg	lac
bif	reg	lac
MIZ	reg	lac
FIBER	reg	tur
NA	reg	tur
SUGAR	reg	tur
tur	reg	tur
des	reg	tur
CA	reg	clo
CAL	reg	clo
FIBER	reg	clo
bif	reg	clo
SATFAT	reg	fus
NA	reg	fus
NA	reg	cau
cau	reg	cau
xan	reg	cau
bif	reg	rhi
rhi	reg	rhi
rho	reg	rhi
ery	reg	rhi
tur	reg	rho
clo	reg	rho
CHOL	reg	bur
FIBER	reg	bur
lac	reg	bur
bur	reg	bur
des	reg	bur
FIBER	reg	des
des	reg	des
MIZ	reg	des
pas	reg	des
FAT	reg	MIZ
FIBER	reg	MIZ
rho	reg	MIZ
SATFAT	reg	cam
bif	reg	cam
fus	reg	cam
des	reg	ent
ent	reg	ent
fus	reg	pas
pse	reg	pas

xan	reg	pas
cau	reg	pse
bur	reg	xan
xan	reg	xan
CA	reg	ery
lac	reg	ery
cau	reg	ery
ery	reg	ery

Appendix D. Human Microbiome MAP-model Equations

act = 17.1947317090088 + 0.227234544300032*p_rho + 0.175778051656537/(p_rhi - 23.4011062795208)

bif = 0.785831107465239*p_bif + 15.2458249927945*FIBER/p_bif

cor = p_cor + 0.139509966963015*p_cau + 3.95343118419879*p_bif/FIBER - 0.000172868152336152*SUGAR*p_cor^2

bac = 78.7500014696244 + 51.7447003873334/(57.8726383953475*p_xan - 1797.33149731475) - 0.164581077019052*p_bif - 0.00209827783966129*FIBER*p_ery

lac = 40.1818558074202 + 0.279073083925966*p_bif + -0.181120215668366*FIBER*p_act/p_MIZ - 0.285913967839153*FAT

tur = 19.8124296826635 + 367.420490954211*p_des*p_tur^2/(NA*SUGAR*FIBER^2)

clo = 43.8461189706724 + 0.316628457923367*FIBER + (CAL + 5.5804007400402*p_bif)/FIBER - 0.0762358440090073*CA

fus = 19.9678852676837 + 0.818683836288815/(80.0246031534957 - SUGAR) + 0.410074822705735/(63.2196727056844 - NA)

cau = 20.0210794455955 + 409.599181569476*p_cau^2/(1086955.39506134*NA + 77338.0397387806*p_xan^2 - 108607.188703335*NA*p_xan)

rhi = 19.9930696210875 + p_rho/(8540.0499990765 - 267.190252916424*p_ery) + (19.4922268060283 - p_rhi)/(p_bif - 80.0398538782457)

rho = 19.999760769633 + 25499.0144717232/(3643590286.02303 + 2094.36414023524*p_tur^2*p_clo^2 - 5525681.47252487*p_tur*p_clo)

bur = 30.9624032752801 + 0.201388089157056*p_bur + -3.02311929602648*FIBER/p_des - 0.00286302363187179*CHOL*p_lac

des = 20.3986236999421 + (0.784402829724915*p_des*p_MIZ - 13.6196917674685*p_MIZ)/(30.4596514370755 + FIBER - p_pas)

MIZ = 20 + 0.475844708415161/(FIBER - FAT) + -0.106979754081751/(p_bur - 36.9778119801856)

cam = 24.7335359486652 + -2.03076219001744/(SATFAT - 20.0375895226518) + p_bif/(p_fus - 79.4028198889855)

ent = -26.0847036297922/(p_des - 80.4432216217961) + -186.201696782273/(0.10809656553104*p_ent - 11.4066396146836)

pas = 2147.94957328786/(108.218823198348 + p_xan - p_fus - p_pse)

pse = 20.6159363029704 + -17.0022020859454/(737.133926507937 - 22.4090977710957*p_cau)

xan = 20.0449206592275 + 0.757484054456456/(p_xan - 39.7120078504478) + 0.0113601408727396/(p_bur - 22.9255035496149)

ery = 32.2041204229574 + 0.429681272546127*p_ery + 0.00707333473907525*p_lac*p_cau - 0.312427420510187*CA

Appendix E. BANJO Parameters for ‘Gradient’ Mouse Microbiome CIN

```
###-------------------------------------------------
### Project information
###-------------------------------------------------
project =                     MinCommunituy
user =                                          PEL
notes =           dynamic bayesian network inference

###-------------------------------------------------
### Input and output locations
###-------------------------------------------------

inputDirectory =             FinalGradient
observationsFile =           Gradeint-init_ForBNI.txt

outputDirectory =            FinalGradient
reportFile =                 BN_Gradient-init.@TS@.txt


###-------------------------------------------------
### Required data
###-------------------------------------------------

variablesAreInRows= yes
variableCount =                                  68

###-------------------------------------------------
### Optional data
###-------------------------------------------------

mustNotBePresentEdgesFile = Gradeint-init_ForBNI_NoInteractions.txt


###-------------------------------------------------
###  Pre-processing options
###-------------------------------------------------

discretizationPolicy =                          i5



###-------------------------------------------------
###  Search specifications
###-------------------------------------------------

searcherChoice =                              Greedy
proposerChoice =                       AllLocalMoves
evaluatorChoice =                            default
deciderChoice =                              default
statisticsChoice =                           default


###-------------------------------------------------
### Search "problem domain" constraints
###-------------------------------------------------

minMarkovLag =                                     0
maxMarkovLag =                                     0
```

```
dbnMandatoryIdentityLags =                         1
equivalentSampleSize =                           1.0
maxParentCount =                                   5
computeInfluenceScores = yes

###-------------------------------------------------
### Search monitoring properties
###-------------------------------------------------

nBestNetworks =                                    5
bestNetworksAre =             nonidenticalThenPruned
screenReportingInterval =                       60 s
fileReportingInterval =                         60 m


###-------------------------------------------------
### Stopping criteria
###-------------------------------------------------

maxTime =                                       12 h
maxProposedNetworks =
maxRestarts =
minNetworksBeforeChecking =                     1000


###-------------------------------------------------
### Parameters used by specific methods
###-------------------------------------------------

### For simulated annealing:
initialTemperature =                            1000
coolingFactor =                                  0.8
maxAcceptedNetworksBeforeCooling =              1000
maxProposedNetworksBeforeCooling =             10000
minAcceptedNetworksBeforeReannealing =           200
reannealingTemperature =                         500

### For greedy:
minProposedNetworksAfterHighScore =             1000
minProposedNetworksBeforeRestart =              3000
maxProposedNetworksBeforeRestart =              5000
restartWithRandomNetwork =                       yes
maxParentCountForRestart =                         3


###-------------------------------------------------
### Misc. options
###-------------------------------------------------

displayMemoryInfo =                              yes
displayStructures =                              yes
```

Appendix F. Mouse Microbiome CIN

```
d_Asp  reg    t_Porphyromonas
d_Calcium     reg    t_Clostridium
d_Calcium     reg    t_Butyrivibrio
d_Glu  reg    t_Atopobium
d_Gly  reg    t_Desulfotomaculum
d_Gly  reg    t_Lactococcus
d_His  reg    t_Oribacterium
d_Ile  reg    t_Collinsella
d_Lactose     reg    t_Erysipelotrichaceae
d_Magnesium   reg    t_Lachnospiraceae
d_Met+Cys     reg    t_Parabacteroides
d_Monosat     reg    t_Eubacterium
d_Starch      reg    t_Lactococcus
d_Starch      reg    t_Lactococcus
d_Starch      reg    t_Clostridium
d_Sucrose     reg    t_OTHER
d_Val  reg    t_Desulfotomaculum
d_Vit A       reg    t_Atopobium
d_Vit D3      reg    t_Erysipelotrichaceae
t_Anaerostipes       reg    t_OTHER
t_Bacteroides reg    t_Ruminococcaceae
t_Butyrivibrio       reg    t_Ruminococcaceae
t_Butyrivibrio       reg    t_OTHER
t_Clostridium reg    t_Lactobacillus
t_Clostridium reg    t_Oribacterium
t_Clostridium reg    t_Collinsella
t_Collinsella reg    t_Ruminococcaceae
t_Desulfotomaculum   reg    t_Bacteroides
t_Lachnospiraceae    reg    t_Ruminococcus
t_Lactococcus reg    t_Desulfotomaculum
t_OTHER       reg    t_Ruminococcus
t_Parabacteroides    reg    t_Clostridium
t_Parabacteroides    reg    t_Erysipelotrichaceae
t_Porphyromonas      reg    t_Clostridium
t_Porphyromonas      reg    t_Atopobium
t_Ruminococcaceae    reg    t_Lachnospiraceae
t_Ruminococcus       reg    t_Clostridium
t_Ruminococcus       reg    t_Clostridium
t_Ruminococcus       reg    t_Clostridium
t-1_Anaerostipes     reg    t_Anaerostipes
t-1_Butyrivibrio     reg    t_Ruminococcus
t-1_Collinsella      reg    t_Ruminococcus
t-1_Erysipelotrichaceae     reg    t_OTHER
t-1_Oribacterium     reg    t_Desulfotomaculum
t-1_Parabacteroides  reg    t_Blautia
t-1_Porphyromonas    reg    t_Clostridiales
t-1_Ruminococcus     reg    t_Clostridiales
d_Ala  reg    t_Anaerostipes
d_Ala  reg    t_Anaerostipes
d_Biotin      reg    t_Desulfotomaculum
d_Biotin      reg    t_Desulfotomaculum
d_Fiber (cellulose)  reg    t_Ruminococcus
d_Fiber (cellulose)  reg    t_Ruminococcus
d_Riboflavin  reg    t_Oribacterium
d_Riboflavin  reg    t_Oribacterium
```

```
d_Ser  reg    t_OTHER
d_Ser  reg    t_OTHER
d_Thr  reg    t_Parabacteroides
d_Thr  reg    t_Parabacteroides
d_Trp  reg    t_Atopobium
d_Trp  reg    t_Atopobium
d_Vit B12     reg    t_OTHER
d_Vit B12     reg    t_OTHER
d_Vit E       reg    t_Lactococcus
d_Vit E       reg    t_Lactococcus
t_Atopobium   reg    t_Lactobacillus
t_Atopobium   reg    t_Lactobacillus
t_Butyrivibrio      reg    t_Erysipelotrichaceae
t_Butyrivibrio      reg    t_Erysipelotrichaceae
t_Clostridiales     reg    t_Butyrivibrio
t_Clostridiales     reg    t_Butyrivibrio
t_Lactococcus reg    t_Oribacterium
t_Lactococcus reg    t_Oribacterium
t_Ruminococcaceae   reg    t_Anaerostipes
t_Ruminococcaceae   reg    t_Anaerostipes
t_Ruminococcus      reg    t_Clostridium
t_Ruminococcus      reg    t_Clostridium
t-1_Lactobacillus   reg    t_OTHER
t-1_Oribacterium    reg    t_Bacteroides
t-1_Oribacterium    reg    t_Bacteroides
t-1_Parabacteroides reg    t_Clostridiales
t-1_Parabacteroides reg    t_Clostridiales
```

VITA

PETER E. LARSEN

Professional Preparation

Purdue University	Ecology, Population Biology	B.S.	1993
University of Illinois at Chicago	Bioengineering	M.S.	2006

Appointments

2009-pres	Assistant Computational Biologist, Biosciences Division Argonne National Laboratory. Systems biology, computational modeling of microbial community interactions,
2001-2009	Biomedical Research Specialist, University of Illinois, Core Genomics Facility Transcriptomic data analysis for biomedical research.
1998-2001	Research and Development, Vysis Inc. Cancer diagnostic microarray development for chip-based CGH.
1997-1998	Scientist, ThermoGen, Inc. Protein engineering for thermostable enzymes in industrial processes
1995-1997	Scientist, FermaLogic, Inc. Metabolomic engineering for enhanced and novel antibiotic biosynthesis.

Publications:

1. S Lax, N Sangwan, D Smith, **P Larsen**, KM Handley, M Richardson, et al. Bacterial colonization and succession in a newly opened hospital. *Science Translational Medicine* 9 (391), eaah6500.
2. S Shinde, JR Cumming, FR Collart, PH Noirot, **PE Larsen**. Pseudomonas fluorescens Transportome Is Linked to Strain-Specific Plant Growth Promotion in Aspen Seedlings under Nutrient Stress. *Frontiers in Plant Science* 8.
3. Luke R Thompson, Gareth J Williams, Mohamed F Haroon, Ahmed Shibl, **Peter Larsen**, Joshua Shorenstein, Rob Knight, Ulrich Stingl (2016). Metagenomic covariation along densely sampled environmental gradients in the Red Sea. *The ISME journal* 11 (1), 138-151 (2016).

4. **Peter E Larsen**. More of an art than a science: Using microbial DNA sequences to compose music. *Journal of microbiology & biology education* 17 (1), 129 (2016).
5. Jessica L. Metcalf, Zhenjiang Zech Xu, Sophie Weiss, Simon Lax, Will Van Treuren, Embriette R. Hyde, Se Jin Song, Amnon Amir, **Peter Larsen**, Naseer Sangwan, Daniel Haarmann, Greg C. Humphrey, Gail Ackermann, Luke R. Thompson, Christian Lauber, Alexander Bibat, Catherine Nicholas, Matthew J. Gebert, Joseph F. Petrosino, Sasha C. Reed, Jack A. Gilbert, Aaron M. Lynne, Sibyl R. Bucheli, David O. Carter, Rob Knight (2015). Microbial community assembly and metabolic function during mammalian corpse decomposition. *Science* DOI: 10.1126/science.aad2646
6. **Peter E Larsen**, Y Dai. Metabolome of human gut microbiome is predictive of host dysbiosis. *GigaScience* 4 (1), 1-16 (2015).
7. **Peter E Larsen**, FR Collart, Y Dai. Predicting ecological roles in the rhizosphere using metabolome and transportome modeling. *PloS One* 10 (9), e0132837 (2015)
8. E Zielazinski, S Zerbs, **P Larsen**, F Collart, PD Laible. Methionine Importers in Soil Bacteria: Potential for Transporter-Component Crosstalk. *Biophysical Journal*. 108 (2), 146a (2015).
9. Simon Lax, Daniel P Smith, Jarrad Hampton-Marcell, Sarah Owens, Kim M. Handley, Nicole Scott, Sean M Gibbons, **Peter Larsen**, Benjamin D Shogan, Sophie Weiss, Jessica L. Metcalf, Luke K. Ursell, Yoshiki Vázquez-Baeza, Will Van Treuren, Nur A. Hasan, Molly K. Gibson, Rita Colwell, Gautam Dantas, Rob Knight, Jack A. Gilbert. Longitudinal analysis of microbial interaction between humans and the indoor environment. 2014. *Science* 345 (6200), 1048-1052.
10. **Peter E Larsen,** Leland J Cseke, R M Miller, Frank R Collart. Modeling forest ecosystem responses to elevated carbon dioxide and ozone using artificial neural networks. *Journal of Theoretical Biology*, Volume 359, 21 October 2014, Pages 61–71.
11. **Peter E. Larsen**, Frank Collart, Yang Dai. Using metabolomic and transportomic modeling and machine learning to identify putative novel therapeutic targets for antibiotic resistant Pseudomonad infections. *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, EMBC 2014 08/2014; 2014:314-7. DOI: 10.1109/EMBC.2014.6943592
12. **Peter E. Larsen**, Dawn Field, Yuki Hamada, Jack A. Gilbert. Satellite remote sensing data can be used to model marine microbial metabolite turnover. ISME Journal. 2014; doi: 10.1038/ismej.2014.107.
13. Yuki Hamada, Jack A. Gilbert, **Peter E. Larsen**, and Madeline J. Norgaard. Toward Linking Aboveground Vegetation Properties and Soil Microbial Communities Using Remote Sensing. Photogrammetric Engineering & Remote Sensing. Vol. 80, No. 4, April 2014, pp. 311–321.
14. Lopez, J., M. Cuvelier, J. A. Gilbert, **P. Larsen**, D. Willoughby, Y. Wu, P. Blackwelder, P. J. Mccarthy, E. Smith, and Vega R. Thurber. "Synergistic Effects of Crude Oil and Corexit Dispersant on a Sponge Holobiont System." In *INTEGRATIVE AND COMPARATIVE BIOLOGY*, vol. 53, pp. E130-E130. JOURNALS DEPT, 2001 EVANS RD, CARY, NC 27513 USA: OXFORD UNIV PRESS INC, 2013.
15. Stephen Lumayag, Caroline E Haldin, Colleen Cowan, Beatrix Kovacs, **Peter Larsen**, Dane P. Witmer, David Valle, Shunbin Xu. Inactivation of the miR-183/96/182 Cluster Results in Syndromic Retinal Degeneration. Accepted at *PNAS* 12/21/2012.

16. **Peter E. Larsen**, Jack A. Gilbert. Microbial Bebop: creating music from complex dynamics in microbial ecology. PLoS ONE, 2013, 8(3): e58119.
17. **Peter E Larsen** and Frank R Collart. BowStrap v1.0: Assigning statistical significance to expressed genes using short-read transcriptome data. BMC Research Notes 2012, 5:275. [**HIGHLY ACCESSED**]
18. **Peter E. Larsen**, Sean M. Gibbons and Jack A. Gilbert. Modeling Microbial Community Structure and Functional Diversity across Time and Space. FEMS Microbiology Letters. Accepted manuscript online: 3 MAY 2012 09:14PM EST | DOI: 10.1111/j.1574-6968.2012.02588.x.
19. **Larsen PE**, Field D, Gilbert JA. Predicting bacterial community assemblages using an artificial neural network approach. Nat Methods. 2012 Apr 15. doi: 10.1038/nmeth.1975. [**Manuscript reviewed in June 2012 Nature Methods, News and Views**]
20. **Larsen P**, Hamada Y, Gilbert J. Modeling microbial communities: Current, developing, and future technologies for predicting microbial community interaction. J Biotechnol. 2012 Mar 23.
21. **Larsen PE**, Collart F, Field D, Meyer F, Keegan KP, Henry CS, McGrath J, Quinn J, Gilbert JA. 2011. Predicted Relative Metabolomic Turnover (PRMT): determining metabolic turnover from a coastal marine metagenomic dataset. Microbial Informatics and Experimentation 2011, 1:4. [**HIGHLY ACCESSED**]
22. **Larsen, Peter E**., Frank Collart, Folker Meyer, and Jack A. Gilbert. "Predicted Relative Metabolomic Turnover-Predicting Changes in the Environmental Metabolome from the Metagenome." In *BIOINFORMATICS*, pp. 337-345. 2011.
23. Havel VE, Wool NK, Ayad D, Downey KM, Wilson CF, **Larsen P**, Djordjevic JT, Panepinto JC. Ccr4 Promotes Resolution of the ER Stress Response during Host Temperature Adaptation in Cryptococcus neoformans. Eukaryot Cell. 2011 May 20.
24. **Peter E Larsen**, Avinash Sreedasyam, Geetika Trivedi, Gopi K Podila, Leland J Cseke and Frank R Collart. Using Next Generation Transcriptome Sequencing to Predict an Ectomycorrhizal Metabolome. BMC Systems Biology (2011), 5:70. [**HIGHLY ACCESSED**]
25. Adler A, Park YD, **Larsen P**, Nagarajan V, Wollenberg K, Qiu J, Myers TG, Williamson PR. A novel specificity protein 1 (SP1)-like gene, regulating protein kinase C-1 (PKc1)-dependent cell-wall integrity and virulence factors in Cryptococcus neoformans. J Biol Chem. 2011 Apr 12.
26. Henry CS, Overbeek R, Xia F, Best AA, Glass E, Gilbert J, **Larsen P**, Edwards R, Disz T, Meyer F, Vonstein V, Dejongh M, Bartels D, Desai N, D'Souza M, Devoid S, Keegan KP, Olson R, Wilke A, Wilkening J, Stevens RL. Connecting genotype to phenotype in the era of high-throughput sequencing. Biochim Biophys Acta. 2011 Mar 21.
27. **Peter Larsen**, Frank Collart and Yang Dai, "Incorporating network topology improves prediction of protein interaction networks from transcriptomic data". International Journal of Knowledge discovery and Bioinformatics, 1(3), pp.1-19. 2010.
28. Park YD, Panepinto J, Shin S, **Larsen P**, Giles S, Williamson PR. Mating pheromone in Cryptococcus neoformans is regulated by a transcriptional/degradative "futile" cycle. J Biol Chem. 2010 Nov 5;285(45):34746-56. Epub 2010 Aug 27.
29. **Peter E Larsen**, Trivedi G, Sreedasyam A, Lu V, Podila GK, Collart FR. Using deep RNA sequencing for the structural annotation of the Laccaria bicolor mycorrhizal transcriptome. PLoS One. 2010 Jul 6;5(7):e9780.

30. **Peter Larsen** and Yang Dai, Using Gene Expression Modeling to Determine Biological Relevance of Putative Regulatory Networks, Proceeding of the 5th International Symposium on Bioinformatics Research and Applications (eds. I. Mandoiu, G. Narasimhan, and Y. Zhang), Lecture Notes in Bioinformatics, Springer Verlag, Vol. 5542 (2009) pp. 40-51, 2009.
31. Kedar Kulkarni, **Peter Larsen** and Andreas A. Linninger, "Assessing chronic liver toxicity based on relative gene expression data", Journal of Theoretical Biology (2008), doi:10.1016/j.jtbi.2008.05.032.
32. **Peter Larsen**, Eyad Almasri, Guanrao Chen and Yang Dai," Incorporating Knowledge of Topology Improves Reconstruction of Interaction Networks from Microarray Data", Lecture Notes in Bioinformatics, Vol. 4983 (eds.by I.I. Mandoiu, Raj Sunderraman, and A. Xelikovsky), Springer Verlag, pp. 434-443, 2008.
33. Eyad Almasri, **Peter Larsen**, Guanrao Chen and Yang Dai, "Incorporating Literature Knowledge in Baysian Network for Inferring Gene Networks with Gene Expression Data", Lecture Notes in Bioinformatics, Vol. 4983 (eds. by I.I. Mandoiu, Raj Sunderraman, and A. Xelikovsky), Springer Verlag, pp. 184-195, 2008.
34. Guanrao Chen, **Peter Larsen**, Eyad Almasri, Yang Dai, "Rank-based edge reconstruction for scale-free genetic regulatory networks", BMC Bioinformatics (2008), 9:75.
35. **Peter Larsen**, Eyad Almasri, Guanrao Chen, Yang Dai, "A statistical method to incorporate biological knowledge for generating testable novel gene regulatory interactions from microarray experiments", BMC Bioinformatics (2007), 8:317. [**HIGHLY ACCESSED**]
36. **Peter Larsen**, E. Almasri, G. Chen and Y. Dai, "Correlated discretized expression score: a method for identifying gene interaction networks from time course microarray expression data" Proceedings of the 28th International Conference of IEEE Engineering in Medicine and Biology Society (EMBS) (2006). pp. 5842-5845.
37. G. Chen, **P. Larsen**, E. Almasri and Y. Dai, "Sample scale-free gene regulatory network using gene ontology", Proceedings of the 28th International Conference of IEEE Engineering in Medicine and Biology Society (EMBS) (2006). pp.5523-5526.
38. Robert Folberg, Zarema Arbieva, Jonas Moses, Amin Hayee, Tone Sandal, ShriHari Kadkol, Amy Lin, Klara Valyi-Nagy, Suman Setty, Lu Leach, Patricia Chevez-Barrios, **Peter Larsen**, Dibyen Majumdar, Jacob Pe'er, Andrew Maniotis. "The generation of vasculogenic mimicry patterns dampens the invasive melanoma cell genotype and phenotype". Am J Pathol (2006), 166:1187-203.
39. Hessler, PE, **PE Larsen**, AI Constantinou, KH Schram, and JM Weber. "Isolation of isoflavones from soy-based fermentations of the erythromycin-producing bacterium Saccharopolyspora erythraea". Appl. Microbiol. Biotechnol. 1997 47(4) P398-404.

Book Chapters

1. **Peter E. Larsen.** Statistical Tools for Study Design: Replication. *Springer Protocols Handbooks*. Humana Press, 10.1007/8623_2015_95. http://dx.doi.org/10.1007/8623_2015_95).
2. **Peter E. Larsen**, Frank R. Collart, Yang Dai. Predicting Bacterial Community Assemblages using an Artificial Neural Network Approach. *Artificial Neural Networks: Methods and Applications*, Springer, New York.

3. Leland J. Cseke, Stan D. Wullschleger, Avinash Shreedasyam, Geetika Trivedi, **Peter Larsen**, Frank Collart. Chapter 12: Carbon Sequestration. *Genomics & Breeding for Climate-Resilient Crops* (ed. Chittaranjan Kole). Springer, New York.
4. Andreas Wilke, **Peter Larsen**, Jack A Gilbert. Chapter 43. *Next Generation Sequencing and the Future of Microbial Metagenomics*. Horizon Scientific Press / Caister Academic Press.
5. **Peter Larsen**, Leland Cseke, Frank R Collart. Using Next Generation Transcriptome Sequencing to Predict an Ectomycorrhizal Metabolome. *Molecular Microbial Ecology of the Rhizosphere*. (Frans J. de Bruijn ed.), INRA/CNRS Laboratory of Plant-Microbe Interactions.
6. Yang Dai, Eyad Almasri, **Peter Larsen**, Guanrao Chen, Structure Learning of Genetic Regulatory Networks Based on Knowledge Derived from Literature and Microarray Gene Expression Measurements, *Computational Methodologies in Gene Regulatory Networks*, (S. Das, D. Caragea, W. H. Hsu, S. M. Welch eds.), IGI Global, pp.289-309, 2009.