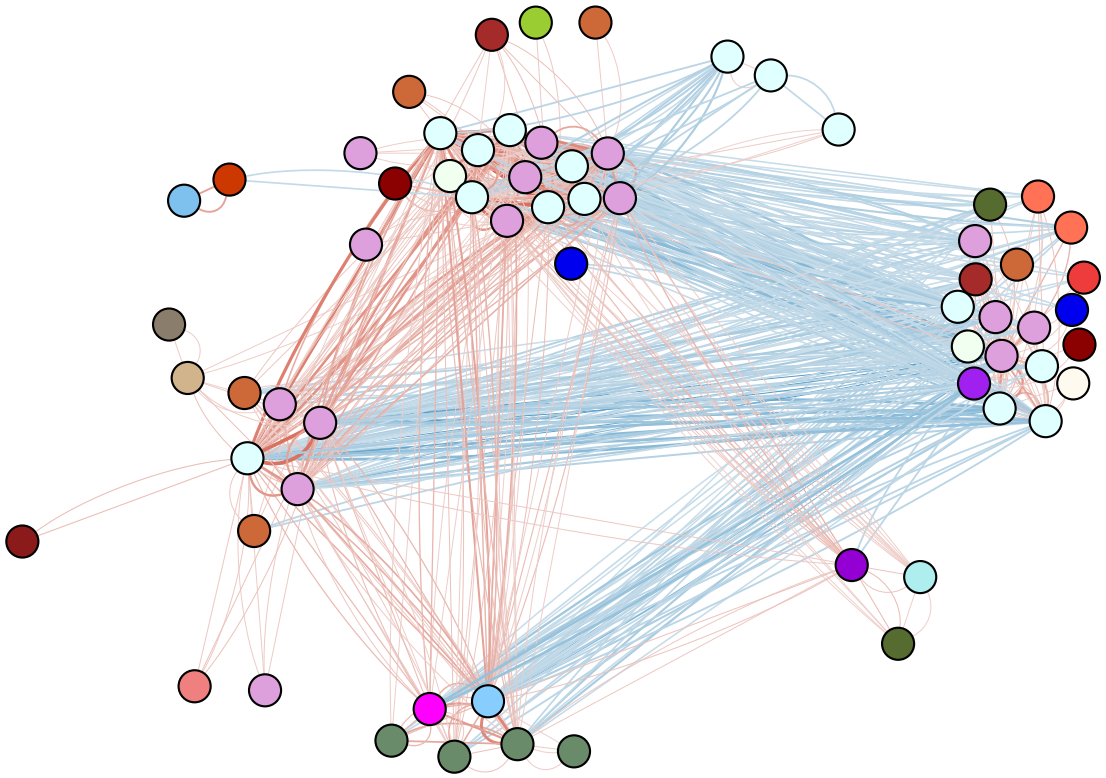
CHALMERS
UNIVERSITY OF TECHNOLOGY



Comparative Network Analysis for Networks Derived from Microbial Compositional Data

with Application to Human Gut Microbiota Disease States

Master's Thesis in Complex Adaptive Systems

WADE ROSKO

Department of Mathematical Sciences
CHALMERS UNIVERSITY OF TECHNOLOGY & KALEIDO BIOSCIENCES
Gothenburg, Sweden & Lexington, Massachusetts, USA 2019

Master's Thesis 2019

Comparative Network Analysis for Networks Derived from Microbial Compositional Data

with Application to Human Gut Microbiota Disease States

WADE ROSKO

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Mathematical Sciences
Chalmers University of Technology
Gothenburg, Sweden 2019

Comparative Network Analysis for Networks Derived from Microbial Compositional Data
with Application to Human Gut Microbiota Disease States
WADE ROSKO

Industrial Supervisor: Stephan Reiling, Fellow in Integrated Data Sciences at Kaleido Biosciences
Academic Supervisor: Rebecka Jörnsten, Department of Mathematical Sciences
Examiner: Rebecka Jörnsten, Department of Mathematical Sciences

Master's Thesis 2019
Department of Mathematical Sciences
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Visualization of the healthy cohort correlation network containing positive (red) and negative (blue) edge correlations (c) meeting the condition: $|c| > 0.15$

Typeset in LaTeX
Gothenburg, Sweden 2019
Boston, Massachusetts USA 2019

Biological Correlation Network Analysis Derived from Compositional Data
with Applications to the Human Gut Microbiota
WADE ROSKO
Department of Mathematical Sciences
Chalmers University of Technology

Abstract

Over the last decade, the human gut microbiota has been identified as a contributing factor to human health. As a result there have been advances in high-throughput technologies that have allowed for increasing amounts of information on the microbiota to be acquired. These technologies have opened the door to a burgeoning field of research in human biology that requires an interdisciplinary approach to better understand the complex relationships within microbial systems and potential interactions with a human host's physiology. One of the most abundant types of data used in the field comes from metagenomics, which is compositional in nature and presents a challenge for identifying key microbes in a microbial community. The aim of this thesis is to generate comparable correlation networks derived from healthy and diseased human gut microbial count data by utilizing more suitable statistical methods. We utilize techniques from compositional data correlation generation methods to create control and case microbial networks. We then apply filtering methods to identify important network connections in the respective systems. From here, we compare the networks using overlapping sub-graphs in order to identify microbes that may influence dysbiosis in the gut microbiota.

Keywords: systems biology, human gut microbiome, correlation networks, community detection, complex networks, compositional data, metagenomics

Acknowledgements

Countless people have influenced my journey as I study and implement methods to learn more about complexity science.

I am extremely grateful for the opportunity to have worked at Kaleido Biosciences this year under the guidance of my industrial advisor and manager Stephan Reiling. Because of his mentorship my scientific and computational thinking has grown considerably. At the same time he and the members on the IDS team have helped show me the relevancy of computational methods in the life science and biotech industries.

At Chalmers I would like to thank Professor Rebecka Jörnsten for working with me as my academic advisor and thesis examiner. Her statistical learning class helped show how much more there is to explore in the life sciences with the use of novel computational methods. Her insight helped me to choose this portion of my microbiome research to present as a thesis. I would also like to thank my thesis opponent, Dimitrios Karypidis, for his friendship and assistance with challenging my thesis defense.

None of this would have been possible without my parents, T.J. Dufresne and T.A. Rosko, who have always done all that they could to foster my thirst for knowledge about the world. It's because of them and their constant support that I have followed this path to explore physics and complexity science. My sister, Becca Jane, has been there through the ups and downs, and I am so thankful to have her by my side offering encouragement and enthusiasm. My grandparents, Dick and Jane, who continuously share their love and belief in me and have helped make this opportunity to study at Chalmers a reality. My grandfather, Stan, who has cheered me on and given me advice as I begin my career.

Finally, I'd like to thank all of my friends that have shared their time with me while I've been in this program. My friends from MPCAS, Övre Fogelbergsgatan 3, Tjugosex, The Hostel, Bates College, Summit, and especially those who have helped me with proofreading and editing.

The work presented in this thesis represents only a portion of the research that I worked on this year. I hope that you find it informative, and are excited about future work in human microbiome research.

Wade Rosko, Boston/Gothenburg, June 2019

Contents

List of Acronyms

AGORA *assembly of gut organisms through reconstruction and analysis* 9
ART arthritis 17, 27
ASD autism spectrum disorder 17, 27

CD Crohn's disease 17, 27
CDI *Clostridium difficile* infection 27
CIRR liver cirrhosis 17, 27
CLI command line interface 29
CLR centered log-ratio 19
CRC colorectal cancer 17, 27

D diseased 45

EDD enteric diarrheal disease 17, 27
ENA European Nucleotide Archive 26, 28

FastSpar Fast Sparse Correlations for Compositional Data . xi, 14, 15, 18, 21, 22, 28, 29, 32, 33, 40
FDN full diseased network ..... xiii, III
FHN full healthy network ..... xiii, III

GRNs gene regulatory networks 4, 7

H healthy 17, 27, 28, 45
HIV human immunodeficiency virus 17, 27
HT high-throughput 1, 5, 6, 8, 10, 49, 51

IBD inflammatory bowel disease 27

JEI Jaccard Edge Index 24, 43

LIV liver disease 17, 27

MHE minimal hepatic encephalopathy 17, 27
mRNA messenger Ribonucleic Acid 7

NASH non-alcoholic steatohepatitis 17, 27
NCBI National Center for Biotechnology Information 26, 28
NESH Neighbor shift ..... xiii, 16, 24, 25, 31, 36, 43–49
NetShift network shift 16, 51

nonCDI control patients with diarrhea who tested negative for CDI 27
nonIBD control patients with gastrointestinal symptoms but no intestinal inflammation 27

OB obesity 17, 27
ODN overlapping diseased network ..... xiii, III
OHN overlapping healthy network ..... xiii, III
OTU Operational taxonomic unit 15, 19, 20, 26, 28, 29, 51

PAR Parkinson's disease 17, 27
PSA psoriatic arthritis 17, 27

RA rheumatoid arthritis 17, 27
RNA ribonucleic acid 8
rRNA ribosomal RNA 9, 19, 26

SparCC Sparse Correlations for Compositional Data ... 14, 15, 18, 19, 22, 29

T1D type I diabetes 17, 27

UC ulcerative colitis 17, 27

WWW World Wide Web 4

List of Figures

List of Tables

1

Introduction

In this chapter, the reader is introduced to the topic of the thesis which includes general concepts, ideas, recent advances, and challenges.

1.1 Background

Biology is the interdisciplinary field of study coupling the knowledge and techniques of the mathematical, physical, and chemical sciences to describe the natural world and the life that inhabits it. Biological knowledge for most of the 20^{th} century was primarily discovered through reductionist techniques, which revealed the observed behavior and structure of cellular components. However, the rise of genomics in the 1990s has significantly changed the biological sciences by opening the door to more paths of inquiry via the increase in available data [Palsson, 2000]. Of note are the high-throughput (HT) technologies being used to generate considerable amounts of data which are becoming more difficult to analyze and understand due to increased size and complexity [Sboner et al., 2011]. With increasing amounts of data available for researchers, there are emerging methods and techniques that can be leveraged to better understand the biological processes underpinning the behaviors and function of biological organisms and molecules.

In this section, the reader will be introduced to the fields of network theory, systems biology, and microbial ecology and their respective ideas and concepts. With these concepts in mind, the reader should have the basic knowledge to understand the scopes and applications of the thesis. Section 1.1.1 covers the network and graph theory and its practical uses in systems research. Section 1.1.2 describes the emergence of systems biology research and the resulting use of *in silico* modeling. Section 1.1.3 introduces the reader to the field of microbiology and microbial ecology and their relevance for study of the human gut microbiota. Together these sections lay the foundation for a systems-based approach to understanding the gut microbiota.

1.1.1 Graphs and Networks

In general, we use the term *graph* or *network* to describe a collection of objects and the information about their relationships between each other. We can use

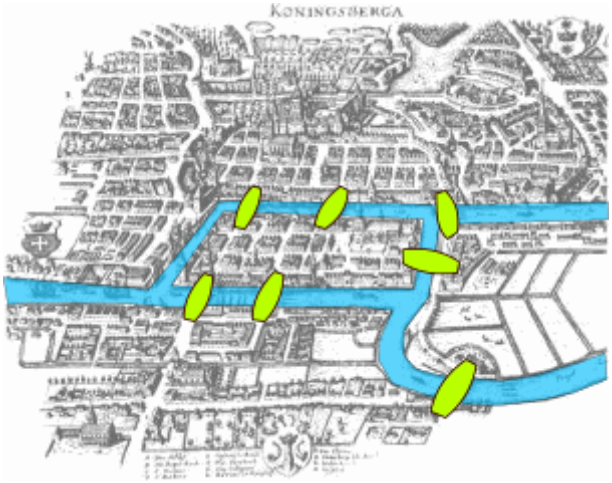


Figure 1.1: An image depicting the Seven Bridges of Königsberg problem. The land masses are defined as nodes, and the bridges (in green) are defined as edges. This image is licensed under the under the CC BY-SA 3.0 license [Giuşcă, 2005].

this method and the abstract mathematical formalism arising from graph theory to describe many real-world phenomena; even those that tend to become more complex. The application of networks in network science spans countless real-world problems and fields of study, thus making it a prominent interdisciplinary field of study.

1.1.1.1 Graph Theory

In 1736, the famed mathematician Leonhard Euler published the first known paper on what would become the foundation of graph theory [Euler, 1736]. His formalism tackled the problem known as the Seven Bridges of Königsberg. The city of Königsberg had three land masses with seven bridges connecting them as shown in Figure 1.1. Euler proved that there was no path to go across all of the bridges only one time each. The modern interpretation of his abstract formalism defines the land masses as nodes (vertices, $\mathcal{V}$), and the bridges as links (E, or edges). His reason for there being no solution was based upon the number of points and the number of connections attached to each point.

It is important to note that the terms *network* and *graph* in graph theory and network science are more-or-less synonymous. Technically, the mathematical description in graph theory describes a graph $\boldsymbol{G}(\mathcal{V}, E)$ that uses the form: {*graph*, *vertex*, *edge*}. On the other hand, network science describes real-world systems which are of the form: {*network*, *node*, *link*} [Barabási and Pósfai, 2016]. Thus, *network* and *graph* and their associated properties will be used interchangeably going forward.

As stated above, a graph $\boldsymbol{G}$ is defined as $\boldsymbol{G}(\mathcal{V}, E)$ and containing the nodes and respective set of edges. Initially, graphs may be split into two different types: *undirected* and *directed* graphs. These describe the directionality of the edge connecting vertex i to vertex j which has a given weight w_{ij}. In *undirected* networks, the edge direction is not important, and the emphasis is placed on the relationship that vertex i is connected to vertex j with a given weight w. In *directed* networks, the order

from i to j is important, and is fundamental to the structure of the network. These networks can either be *weighted* or *unweighted*. The w_{ij} values in *weighted* networks vary based upon the weights associated with the connection, whereas the w_{ij} values in a *unweighted* network are set to 1.

A graph may be depicted in one of three ways: 1) A **square adjacency matrix** A, which depicts all of the possible edges connecting vertex i to vertex j with a given weight w_{ij} (so $A_{ij} = w_{ij}$). In this case, a directed network may be asymmetrical due to different edges connecting nodes with varying weights. An undirected network will always be symmetrical since the path from i to j and j to i are equivalent. 2) An **edge list** listing all of the nonzero edges in the network. This method usually describes the graph in a list where each entry takes the form: {node i, node j, w_{ij}}. It is commonly used in computational approaches because a majority of graphs are sparse and not fully connected, so edge lists reduce complexity in calculations by removing arbitrary information. 3) A **visual graphical representation** where nodes are points in the graph, and edges are depicted as lines between nodes. The graphical representation is up to the individual to change, but common implementations will change node sizes based on the number of edges connecting them, and modify edge width based upon the value of the weight.

1.1.1.2 Network Science

Network science emerged as a new field of study by applying basic principles of graph theory to real-world phenomena. As mentioned in Section 1.1.1.1, networks are graph-based systems that describe interacting agents in the everyday world. As a field of study, network science is relatively new – emerging in mainstream science towards the end of the 1990s.

Prior to this in the late 1950s, Erdös and Rényi introduced a paper on *random graphs*, which describe networks whose nodes have probabilities p of connecting to other nodes in a network [Erdös and Rényi, 1959]. Networks at this point were regarded as "regular" with uninteresting topology and features, or very "complex" and "random" as associated with random graphs [Vespignani, 2018]. It was not until 1998 when Watts and Strogatz proposed a new model called the *small-world* network [Watts and Strogatz, 1998] which defined a middle ground between a regular and random network. This model was better at describing the small average *path length*[1] of real-world networks and nodes that were more connected than others. Interestingly enough, the next paper to fully push network science into many domains was when Barabási and Albert published their papers in 1999 on preferential-attachment and power laws [Barabási and Albert, 1999, Barabási et al., 1999, Albert et al., 1999]. The degree distributions of many systems follow power-law distributions, and are "heavy-tailed" compared to the normal distribution of random graphs.

From here, specialists across all types of disciplines started to implement network

[1] Path length is defined as the shortest path between two nodes in a network. Average path length l_G is defined as: $l_G = \frac{1}{n(n-1)} \sum_{i \neq j} d(v_i, v_j)$ where $d(v_i, v_j)$ is the distance between vertices v_i and v_j.

science based analyses in their respective field of study. For example, some of these fields range between sociology (social network analyses), biological systems (epidemiological models, gene regulatory networks (GRNs), etc.), information-technology networks (World Wide Web (WWW), computers), and countless other disciplines. Barabási expresses the interdisciplinary nature of network science quite well:

> "Today many fields consider network science their own. Mathematicians rightly claim ownership and priority through graph theory; the exploration of social networks by sociologists goes back decades; physics lent the universality concept and infused many analytical tools that are now unavoidable in the study of networks; biology invested hundreds of millions of dollars into mapping subcellular networks; computer science offered an algorithmic perspective, allowing us to explore very large networks; engineering invested considerable efforts into the exploration of infrastructural networks. It is remarkable how these many disparate pieces managed to fit together, giving birth to a new discipline."
>
> – Barabási and Pósfai [2016]

Appreciating the scope of network science research presented in the previous quote is helpful when approaching a novel problem because it may be solved by using a technique developed in one of the many different applicable fields.

1.1.1.3 Complex Systems

In the last couple of decades, the growth of network science also led to the growth and study of complex systems. *Complex system* is a name generally given to agents, organisms, structures, and other phenomena that exhibit emergent behaviors and properties [Foote, 2007]. Such objects and systems tend to be complicated (as the name suggests), as they often exhibit chaotic and non-linear interactions. A term most often associated with complex systems is *emergence*, which describes the unexpected outcomes from the system.

Emergence arises because the properties of an individual agent in a system may be fully known. However, as the interactions between multiple interacting agents in a network are examined, it is often not possible to study and predict the exact outcomes from each interaction. Such behaviors are found frequently in real-life; complexity science's role has been to understand these instances with both established and novel theoretical techniques. One of the most important characteristics of complex systems is that interaction types are not limited. These could be anything from an exchange of energy, momentum, material or information [Werner, 1999]. The varying types of interactions that can be modeled allows for most experimental and some theoretical problems to be explored and further researched. Note that emergent behaviors in complex systems are often *robust* – meaning that even if some microscopic interactions are rewritten, the ultimate behavior of the system is similar despite a change in the initial parameters. Robustness is a fundamental property of some complex systems because if they were not robust, truly chaotic behavior would

emerge[2].

Complex systems are often described by network science and graph theory methods since they are able to be described by networks, but the tools that are used to investigate them vary quite largely. Some common tools used to describe and research the systems take a bottom-up approach or a top-down approach [M'hamdi and Nemiche, 2018].

A typical type of bottom-up approach would be agent-based modeling, where individual agents are defined and given certain probabilistic characteristics and behaviors. Then their interactions within the model can be modeled based upon the unique agent's behavior while interesting metrics are recorded as the simulation runs. Another bottom-up approach would be the use of cellular automata as they are similar to agent-based models where the immediate surroundings of an individual impact its behavior. Together these microscopic interactions turn into some type of macroscopic behavior that we would define as an emergent behavior.

Top-down approaches look at known causal mechanisms and behaviors of individual interactions that can be represented via numerical and analytical nonlinear dynamical systems. In the case of an analytical description of the system, we would expect the result to differ drastically from the complex system's actual behavior since it would generalize the inherently large parameter space. Using numerical methods would allow for the parameters of the system to be better identified because the dynamics might not have analytical solutions or be described with a close-form solution. The top-down approach is a more holistic approach since we try to generalize the behavior of the system through guiding equations.

1.1.2 Systems Biology

As mentioned previously, biology research has focused upon gaining new knowledge through reductionist techniques, but novel research into HT technologies has increased the scope of biological research. Today, we find that biological research has extended to a more integrative approach that includes the use of integrative analysis, bioinformatics, mathematical models, and *in silico* (computer simulation) models [Palsson, 2006]. Together this is termed as systems biology, which extends beyond the component-based biology that reductionist techniques are derived from.

1.1.2.1 Network Structure and Emergent Functionalities

Since we have growing lists of information on cellular components derived from HT technologies and we know that cells, biological molecules and structures lead to overall behaviors, with inductive reasoning we assume that there must be behaviors at the microscopic level that lead to macroscopic behaviors. Thus, we want to figure out how emergent cell behavior and functionality arise. A systems perspective allows

[2]One example supporting this is a topic discussed in this thesis in Section 1.1.4. We discuss how a microbial community can have varying distributions of microbes at a genus or species level, but the overall function of the group remains the same.

us to first use the information and properties of our cellular components to then model interactions that lead to emergent behaviors and reveal functional roles of specific groups. Figure 1.2 sketches out the difference between the 20th and 21st-Century approach to biology, and shows that the integrative approach of systems biology uses different methods that utilize the information provided by components biology.

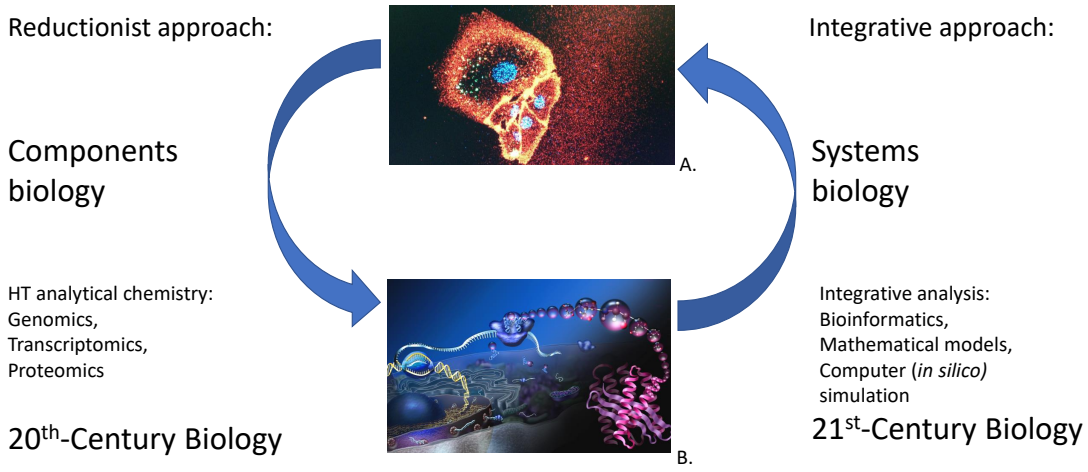


Figure 1.2: Diagram indicating the 20th-Century approach of reductionist biology, where biology is broken down into components. In this case we depict human cells **A**, and an artist's depiction of different omics functions **B**. With HT technologies (such as the omics methods listed) emerging in the 20th-Century, we have acquired the information necessary to know what makes up biological organisms. This has led to the integrative approach of systems biology and the work that is presently being performed today in the 21st-Century. The diagram was redrawn from Palsson [2000, 2006] and the subfigures **A** and **B** are available from Vidal [2013] and Rager [2012] respectively under Creative Commons licenses listed in their citations.

Cellular functions related to the combined behavior of a group of system structures (genes in this case) make up the components that lead to system-level understanding. These functions are often defined as *genetic circuits*, *cellular wiring diagrams*, and modules [Palsson, 2006]. We can build frameworks for modeling specific intracellular behaviors through genetic circuits, and then we can describe physiological behaviors as emergent functions of multiple genetic circuits. This type of process allows the systems approach to work because we can map function between the different circuit components. Utilizing this component-based approach means that the information produced by HT technologies allow for the genome to actually be the system we are investigating and modeling.

Overall, a system-level understanding of biological systems is dependent upon four properties: system structures, system dynamics, control methods, and design meth-

ods [Kitano, 2002]. System structures are found via our components biology methods and represent gene interactions, biochemical pathways, and physical properties of biological structures. Insight into system structures today have been heavily dependent on laboratory experimentation, expression profiling, and transcription regulation analyses of messenger Ribonucleic Acid (mRNA). System dynamics knowledge requires the construction of models based upon information related to system structure. In turn, model building requires the scope of the model to be defined prior. These are extremely complicated systems containing many parts, so the scope of the model requires a focus and a resolution level. Most models take similar forms to the approaches described in Section 1.1.1.3. To learn about the system as a whole, control and design methods are used to modulate biological system states through different mechanisms [Kitano, 2002]. Together these factors allow for different levels of a biological system's hierarchy to be linked.

1.1.2.2 Computational Biology as it Relates to Systems Biology

Computational methods for modeling biological behavior have been around for nearly 60 years. Some of the first biological simulations were run on analog computers by Goodwin [1963]. Goodwin modeled the oscillatory behavior in GRNs by exploring self-negative feedback loops for a gene that codes for the production of a metabolite that in turn inhibits the expression of the gene itself. This initial modeling led to many other studies using computers between the 1970s and the 1990s to simulate large metabolic networks, cell-scale models, genome-scale models of viruses, genome-scale metabolic models of bacteria, and large-scale models of mitosis [Palsson, 2006]. The past twenty years have seen a significant increase in the scope of computational biology work, and most of this is due to the shift in computational capabilities as well as the ability to record more information from experiments and explore biology significantly better at the genome level. An interesting opinion from 2002 stated that biology is set to become a quantitative heavy science because the need for numerical analysis and modeling to discover system-wide analytical theories is necessary. Noble [2002] argues that qualitative thinking fails with the complexity of biological systems and is under the opinion that biology would become one of the most computer-intensive sciences this century. Observing the field today, it would be fair to say that they have been right so far about the role of computation in biology. Most academic and industrial biology research requires the aid of individuals well-versed in computational methods, for they help make sense of data from all hierarchy levels of the systems being investigated, and they play a large role in guiding research questions that are based on analyses and computational models.

1.1.3 Microbiology

Despite existing for billions of years and being some of the most abundant and diverse life forms on the planet, human research of microorganisms only started in the late 17th century. In 1676, Dutch scientist Antonie van Leeuwenhoek successfully created the first single-lensed microscopes and wrote on the presence of protists and

bacteria living in different environments [Lane, 2015]. His research created the field of microbiology which focuses on the study of biological entities that are too small to be seen by the unaided eye. These entities may include individuals from the Archaea, Bacteria, and Eukarya domains, and more specifically include bacteria, archaea, protists, fungi, parasites, and viruses [Sattley and Madigan, 2015]. Considering that microorganisms are so abundant and represent such a diverse population of organisms, this thesis focuses on the ecological communities that microorganisms live in.

1.1.3.1 Microbes, Microbiota and the Microbiome

Microorganisms can be found all over the earth inhabiting mundane and extreme places alike: from the upper parts of the atmosphere [Fulton, 1966] to high temperature and pressure environments of submarine hydrothermal vents [Anderson et al., 2011]. In all of these environments microbial[3] communities emerge and flourish as they compete and support each others' metabolic systems. With new HT technologies, the ability to study individual microbes and microbial systems has improved, and the application of systems biology techniques promises novel discoveries in the realm of microbial ecology.

To the casual observer the terms *micobiota* and *microbiome* appear to be interchangeable, but they have subtle differences. The term *microbiota* is commonly used to describe the group of microorganisms in an environment [Marchesi and Ravel, 2015]. With this term we can discuss communities of microorganisms as microbiotic systems: e.g. the human gut microbiota, soil microbiota, plant microbiota, etc. We use *microbiome* to describe the entire habitat including the microorganisms, their genomes, and environmental conditions [Marchesi and Ravel, 2015]. the description of the microbiome can be compiled with and of the multiple types of -*omics* data generated using HT technologies: *metagenomics*, *metabolomics*, *metabonomics*, *metatranscriptomics*, and *metaproteomics*. Metagenomics focuses on the data related to gene sequences that are found in a sample which are used to identify taxa. Metabolomics refers to the data describing metabolites that are produced by a strain or sample. Metabonomics describes the metabolite data that is the product of multiple strains of organisms. Metatranscriptomics is the analysis of the data expressed by ribonucleic acid (RNA) which is genomic information. Metaproteomics represents the data containing the protein profiles in a sample. Of all of the -*omics* data, metagenomics is used most frequently to identify community profiles of different taxa in varying microbiota [Knight et al., 2018]. The technologies used to generate

1.1.3.2 Microbial Interactions

In Sections 1.1.1 and 1.1.2 we discussed the relevance of a systems-level approach to studying real-world interactions; these easily extend to systems research applied to

[3]Even though *microbe* and *microbial* are usually associated with bacteria, *microbe* is often used interchangeably with *microorganism*.

microorganisms. One example to validate this extension is to consider that a recent estimate predicts that there is a 1:1 ratio of microorganisms to human cells in the human microbiota [Sender et al., 2016]. If this estimate is valid, then that places the number of microbes in a typical adult human around 10 trillion. If interactions between agents scale considerably as a network increases, then a network of these microorganisms and their interactions with each other and their environment will certainly be complicated. Additionally, the complexity of microbial interactions is not limited to massive systems. If we reduce a microbiome down to include only a few different types of microbes, then the behavior of the microbiome could still develop emergent properties and complexity. There is currently a significant amount of research in the field of microbiology that is focusing on microbiome interactions through a network lens [Layeghifard et al., 2017].

As one of the most frequently used and accessible -*omics* data types, genomics allows for mapping of the composition of communities with a high resolution. These maps have increased interest into ecological mechanisms that govern microbial communities and emergent functions [Costello et al., 2012]. In the field of microbiology, metagenomics most frequently targets the rRNA of microbes. Therefore a quantitative and predictive understanding of microbiome ecology is of interest so that we may acquire a better understanding of functions and can use the knowledge to manipulate various microbial systems [Goldford et al., 2018].

A key ecological related system interaction is the competition for resources between individuals in the environment. This is relevant for microbial networks because resources are a primary driver of proliferation. For the most part metabolites are the resources in microbial networks, and they can impact the microbes in multiple ways. Microbes may consume metabolites, which in turn allows them to reproduce and excrete fermented metabolites. Other metabolites may limit a certain microbe's growth. With genomic information and wet-lab experimentation we can identify genes that target different metabolites, and we can begin to identify how microbes are affected by metabolites. This metabolic flux through the system gives a finer resolution of the microbial ecological network, but it is quite difficult to isolate and map the true flux through complicated systems without robust temporal metabolomic and transcriptomic data. One way to help with this mapping is to generalize interactions between microbes, and create genome-scale metabolic reconstructions of microbes present in the environment being studied.

A recent, notable effort along these lines is the *assembly of gut organisms through reconstruction and analysis* (AGORA) framework developed by Magnúsdóttir et al. [2016]. In the framework, the authors have reconstructed 773 microbes present in the human gut microbiota[4]. The metabolic constructions can be used to then model metabolic flux and microbial populations, and ultimately allow for the investigation of inter-species interactions. In the paper the authors use AGORA to model the metabolic flux and microbial populations based upon different diets. They then extrapolated the pairwise microbe-microbe interactions into 6 different interaction

[4]Magnúsdóttir et al. are still updating AGORA with additional microbes. Today there are 812 microbes in the database. The AGORA reconstructions are available via the Virtual Metabolic Human web tool: `https://www.vmh.life/#home`.

types [Magnúsdóttir et al., 2016] based upon co-growth. The interaction types are competition, parasitism, amensalism, neutralism, commensalism, and mutualism and all have been found by comparing the co-growth rates of the microbes with the growth rates of the microbes if they were grown individually. Competitive interactions are classified when the growth rates of the two microbes are both significantly less than the growth rate if the individuals were grown on their own in a medium. Parasitism sees increased growth of one microbe, while the other has a significant decrease in growth rate when grown together. Amensalism occurs when one microbe remains unaffected while the other's growth significantly decreases. Neutralism is when co-growth rates remain unchanged from the individual growth. Commensalism is similar to amensalism, but instead of one microbe's growth decreasing in co-growth, it increases while the other remains the same. The final interaction classification is mutualism where both microbes have increased growth rates in co-growth as opposed to being grown individually. These interactions are clearly defined for pairwise interactions, but multiple level interaction techniques are still being developed.

As HT technologies advance, the ability to aggregate more robust frameworks will be possible, but modeling then proves to be a computational and mathematical problem when scaling up metabolic interactions with many microbes.

1.1.4 Community Structure

Microbial ecology has heavily influenced and contributed to current understanding of microbial system structure and function, but emergent structure is still not fully understood. As ecological systems, microbiotas have the ability to be modeled via resource flux, agent-agent interactions, and population dynamics. At the agent interaction level, such environments tend to reveal emergent structure. Structure can be resolved by tracking community composition and identifying clusters of microbial interactions. Research suggests that community structure is stochastic in nature [Robinson et al., 2010, Nemergut et al., 2013, Zhou et al., 2013], but recent findings argue that structure is conserved at various taxonomic levels [Goldford et al., 2018].

Goldford et al. [2018] managed to observe various community distributions at varying taxonomic levels. In their research, they investigate community stability attractors by populating a sample with various bacteria strains and a given resource. After letting the bacterial communities grow until community stability was reached, they coined the term *guild* to describe the identified groups within the stable communities. Across many experiments and samples the family level distributions of the communities and guilds were conserved, but genus and species level distributions were found to vary between experiments. The authors concluded that the guilds formed at the family level have certain functionalities in the community's overall stability state, and that microbes that were swapped at the genus and species level must exhibit similar roles in the guild if the overall family and community behavior is similar.

1.1.4.1 Microbiotas and Hosts

While microbiotas are found all over the world, some of the most pertinent types are those that coexist with other organisms. Most living animals and plants have developed mutualistic relationships with microorganisms to extend their metabolic capabilities by evolving specialized organs to assist in nutrient acquisition [Hacquard et al., 2015]. Genomes of hosts encode for different enzymes, and while a host may have enzymes that break down certain types of molecules, there are still a large number that the host may not be able to use on its own. The co-evolved system of the host and its microbiota allows for additional types of small molecules to be broken down with enzymes unique to the microbiota. This behavior underlies one of the central mechanisms of host-microbiota behavior; the relationship between a host and its microbiota is significantly dependent upon the metabolite exchange between the two. The field has been trying to identify differences in microbiota states and research indicates that microbiotas influence host health and microbiome composition, especially in microbiotas where there is a competition for resources [Hacquard et al., 2015].

The human gut microbiota's composition and stability has been associated with various types of diseases ranging between metabolic and neurological to cardiovascular and autoimmune disorders. While there are studies showing the associations of the diseases with the gut microbiome, we do not know specifically which types of microbiota states contribute to a disease and which diseases impact the gut microbiota. However, newer asserts that it is possible to identify a healthy or diseased person from their microbiome data [Duvallet et al., 2017]. In the coming years it is expected that the directionality of the associations will be further understood as well as the influence of additional variables on a host (e.g. diet, environmental conditions, medical history, prescriptions, etc.).

1.2 Analyzing Correlation Networks to Identify Differences in Comparable Networks

With recent advances in knowledge of the human microbiota, the gut microbiome has been identified as having strong associations with human health. Since a significant amount of research has identified differences in healthy and diseased individuals' guts, we know that these microbial communities differ between cohorts. Gut microbiome knowledge may be useful in future treatments or diagnoses of diseases, so a need arises to better understand microbial interaction networks. This thesis aims to make use of several different technologies and methods in order to help identify microbes that may be central components of diseased networks. With this knowledge, researchers in the future may be able to use it to help in the modulation of gut microbiotas.

This thesis utilizes compositional-based statistical methods to generate microbial correlation networks derived from compositional data. Since current knowledge of

the human gut microbiota is derived from compositional data, we require formal methods to deal with compositional data that may contain both highly dominant and rare taxa. And since this compositional data may vary greatly between reads in the same sample or across labs, correlation generation methods must be utilized to allow for a comparison to be formulated.

In the thesis we use human gut microbiota to compare correlation networks from control (healthy) and case (diseased) cohorts. Prior to comparing the correlation networks we need to filter out noisy features so that important signals remain. There are no clearly defined methods for network filtering so a two-step arbitrary filtering method is employed to reduce potential artifacts in the overlapping network sub-graphs. Using a recent technique, we analyze the overlapping sub-graphs to identify microbes that may drive a healthy system to a diseased one.

To identify driving microbes, we use several different techniques stemming from a range of compositional data and microbiome research. In the following chapter we briefly go over the key papers and techniques allowing us to take compositional metagenomic data and identify driving members of a network.

2

Related Literature

This chapter reviews some relevant literature and techniques that are referenced and utilized in this thesis. First, the basis behind guilds and the relevance to the human gut microbiome is discussed [Goldford et al., 2018]. Then the SparCC [Friedman and Alm, 2012] and FastSpar [Watts et al., 2018] algorithms will be introduced, followed by the case and control network comparison technique from NetShift [Kuntal et al., 2018]. Additionally we mention the MicrobiomeHD project [Duvallet et al., 2017] which contains the data used in this study.

2.1 Emergent simplicity in microbial community assembly

As mentioned in the Background section, Goldford et al.'s research revealed that there is conservation of microbial distributions across the family level in systems that have been allowed to grow to a stable state. Goldford et al.'s motivation was to investigate whether taxonomic architectures can be explained and predicted by fundamental quantitative principles. Figure 2.1 contains plots showing the composition at the species and family levels across 20 experiments with three different media present in the growth environments.

Despite having the same initial populations and resource amounts, species level compositions at the stability point varied considerably, but the family level compositions remain conserved across studies. From these experiments and multiple other analyses, the authors determined that resources available to microbes significantly dictated the structure of a stable community. This type of behavior is not unique to the experiments, but can be extended to various other microbiomes such as the human gut, plant foliage, and oceans [Goldford et al., 2018]. In general, the authors' work suggests that there are mechanisms that contribute to community-wide stability states. Part of this indicates that stabilized microbial communities consist of metabolic generalists instead of specialists, meaning that members of the communities carry certain functional roles. As they are swapped out for different microbes with similar roles, the overall behavior and family-level stability stays static.

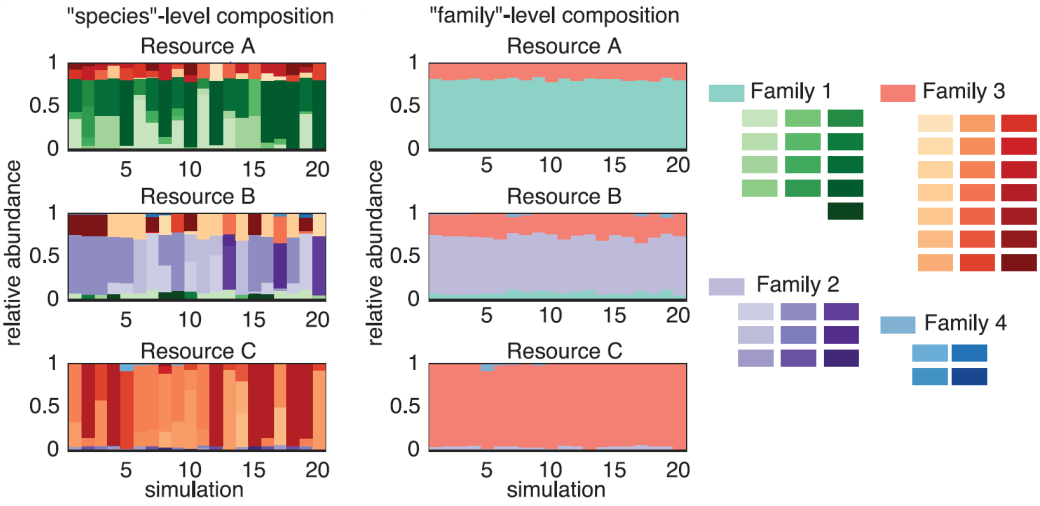


Figure 2.1: Community distribution plots from Goldford et al. The plots on the left look at the community composition at the species-level, and those on the right show the family-level composition. Family level composition is conserved across simulations and resource types. This image from [Goldford et al., 2018] has been licensed for reuse in this thesis with permission from Science and The American Association for the Advancement of Science under the license number 4605400364700.

2.2 Sparse Correlations for Compositional Data

The Sparse Correlations for Compositional Data (SparCC) and Fast Sparse Correlations for Compositional Data (FastSpar) methods are both very similar, but the latter has an improved p-value estimation. Together the techniques aim to take compositional data and treat counts as relative amounts.

2.2.1 Inferring Correlation Networks from Genomic Survey Data

This paper introduces the SparCC method which was proposed by Friedman and Alm [2012]. It was developed because genomic data gives relative, not absolute, abundances of the components of a community. The fact that the data is relative stems from the variabilities in sequence read depth and length. Metagenomics sequencing aims to resolve a general understanding of the taxonomic structure of a sample based upon high quality sequence reads. Therefore, absolute counts are not a given because many sequences are often filtered out based on quality. Common methods in the literature treat the compositional data as absolute and normalize observed counts by the total number of counts to get the fractional abundances. The problem here is that correlation estimates are biased since the fractions are not independent [Buccianti et al., 2006] and the resulting correlations are based upon fractional relationships instead of the underlying biological processes [Aitchison,

2003, Friedman and Alm, 2012].

Friedman and Alm's analysis of standard correlation inference techniques concluded that it performs poorly on genomic data by looking at how the fractional abundance variation of one taxon greatly influences all of the other taxa in the opposite direction. This creates artificial positive and negative correlations based upon abundance values. To circumvent this the authors assume that the number of different taxa is large, and that the resulting networks are sparse. Their method takes the original abundances as basis variables and then assigns a fractional abundance based upon a sampling of a Dirichilet distribution of the basis variables. They then estimate the linear Pearson correlations between log-ratio transformed components. Friedman and Alm determined that their method is robust to violations of the sparsity assumption, does not unfairly treat rare taxa, is highly accurate on simulated data, and identifies phylogenetically structured correlations [Friedman and Alm, 2012].

The resulting correlation and statistical significance matrices represent fully connected networks that contain correlation connections in the range $[-1, 1]$ for each possible OTU pair. In the literature, authors tend to employ some type of threshold value to filter out unnecessary information and connections that may not be significant. This is most frequently done by modifying threshold values to obtain a network that is scale-free or meets some type of average path length or centrality measure, but even this is performed arbitrarily [Perkins and Langston, 2009, Batushansky et al., 2016, Romero-Campero et al., 2016]. Following arbitrary threshold selection, Friedman and Alm employ arbitrary thresholds when pruning the resulting networks to analyze. They selected edges greater than 0.3, and omitted unconnected nodes. Arbitrary threshold selection is neither right nor wrong, and the prevalence of truly scale-free network structures is a controversial topic in network science today [Broido and Clauset, 2019, Barabási, 2018, Holme, 2019].

2.2.2 FastSpar: rapid and scalable correlation estimation for compositional data

The FastSpar method developed by Watts et al. [2018] takes the SparCC method of Friedman and Alm and implements the algorithm in C++. The SparCC program requires a great deal of memory and compute time for high dimensional data sets. The FastSpar implementation was designed to decrease runtime of the SparCC algorithm and to implement a less biased p-value calculation. Based upon Phipson and Smyth's paper, Watts et al. argue that the p-value estimator in SparCC is biased. To combat this they use a p-value estimation that corrects p-value understatement by allowing for the possibility that permutations of the data are repeated [Watts et al., 2018, Phipson and Smyth, 2010]. According to their results, FastSpar was up to 821× faster with a memory reduction of up to 116× less than SparCC using 16 threads, and also generated more accurate p-value estimations.

2.3 'NetShift': a methodology for understanding 'driver microbes' from healthy and disease microbiome datasets

The NetShift (network shift) methodology is proposed by Kuntal et al. [2018] to quantify the rewiring and changes in two different microbiome networks (defined as the case and control networks). The authors developed the NetShift method so that they could generate biologically-backed inferences of microbiome interactions based upon the microbial interaction networks of the respective microbiomes. In order to perform a reasonable comparison between networks, Kuntal et al. search for the shared nodes between the networks and extract all of the edges associated with them. The resulting sub-graphs are then analyzed to identify community-level changes which can identify key microbes in the networks, or network structure shift as defined by the authors.

Kuntal et al. quantify the overall changes in the networks by first looking at the Jaccard edge index (unique edges in a network compared to all of the edges in the two networks) and then running a hierarchical clustering algorithm to label communities based upon the network structure. They run the clustering on both the control and case networks, and then visualize the re-wiring of the network by tracking the community that each node is associated with in each clustering network. With this knowledge, the authors argue that one can identify significant community changes and identify specific microbes that are taking part in the community shuffling. While not referenced in the paper, this idea could be associated with the work of Goldford et al. [2018] and their "guilds". Kuntal et al. then use a new metric to evaluate the changes in associations of a single node in the network which they call Neighbor shift (NESH). NESH is formally defined in Section 3.2.2 and evaluates the neighborhood associations of each node. Nodes with high NESH and betweenness are then considered to be "drivers" of the network.

With the NetShift methodology we find microbes that may contribute to different community functionality and we can gain a better idea of the stability of the networks' structure between states.

2.4 Meta-analysis of gut microbiome studies identifies disease-specific and shared responses

Microbiome research has produced hundreds of studies linking human health and the human microbiome, but few try to leverage the statistical power of aggregating this data. Duvallet et al. [2017] performed a meta-analysis using 28 different studies which all contained control and case cohorts covering 10 different diseases. The authors were able to create a random forest model to classify healthy and diseased individuals solely using their microbiome sample. Such results suggest that there is inherent structure in the communities of healthy and diseased individuals regardless

of the type of disease.

The meta-analysis included patients with the following health-disease states: arthritis (ART), autism spectrum disorder (ASD), Crohn's disease (CD), liver cirrhosis (CIRR), colorectal cancer (CRC), enteric diarrheal disease (EDD), healthy (H), human immunodeficiency virus (HIV), liver disease (LIV), minimal hepatic encephalopathy (MHE), non-alcoholic steatohepatitis (NASH), obesity (OB), Parkinson's disease (PAR), psoriatic arthritis (PSA), rheumatoid arthritis (RA), type I diabetes (T1D), and ulcerative colitis (UC). Together these represent various types of diseases – metabolic, neurological, cardiovascular, autoimmune, and other disease types. The binary classifier implemented by Duvallet et al. reveal that health state is related to dysbiosis in the gut. Given this data, the authors also tried to build a multi-class classification model. The results were poor – resulting in low accuracy measures. The failure to build a model for these disease types is probably attributed to the imbalance in classes and lack of sufficient data points to distinguish specific diseases from each other.

The study by Duvallet et al. [2017] is an important first step in drawing conclusions from publicly available data. Going forward, the authors hope that the microbiome field will continue to make its data and associated patient metadata publicly available. A large concern of this study was batch effect, so future analyses should include additional disease types and data for the included studies in order to minimize potential batch effects. With additional data there should also be a better representation of possible microbiome states associated with the various diseases.

3

Theory

This thesis utilizes techniques for interpreting correlations from compositional data from two different cohorts. The statistics based methods of SparCC and FastSpar (Section 2.2.1) transform the compositional data into the two respective correlation networks. Then, overlapping sub-graphs are generated between the two networks which allows for the networks to be compared and for the identification of ”driver“ nodes to be identified. This comes from the general methodology from NetShift (2.3).

3.1 SparCC and FastSpar

SparCC utilizes the log-ratio transformation to create correlation networks from compositional data. This is a standard technique used in compositional data in order to reduce compositional effects and the possible impact that rare taxa may have in a correlation analysis. Using the log-transformed components of the data, SparCC estimates the Pearson correlations to an approximation based upon the assumptions that the number of components is large, and the correlation network is actually sparse [Friedman and Alm, 2012]. FastSpar follows the same methods as the SparCC algorithm, but differs in the implementation of assessing the statistical significance. Both of these

3.1.1 Bayesian Approach to Estimate Component Fraction

Initial normalization of the compositional data is required for later stages in the SparCC algorithm, however the maximum-likelihood estimate of normalizing by the total counts in the sample is unreliable. This method overestimates the total number of zero fractions from rare taxa or variations in sequencing depth [Agresti and B. Hitchcock, 2005, Friedman and Alm, 2012]. Instead, the authors estimate the component fractions with a Bayesian approach by utilizing the Dirichlet distribution.

The SparCC approach utilizes the full joint distribution of fractions generated by the Dirichlet distribution. Generally the Dirichlet distribution is defined as $\mathrm{Dir}(\alpha)$ where α is a parametric vector of positive reals. Using the Gamma distribution:

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt, \tag{3.1}$$

the Dirichlet probability density function can be formally defined as:

$$\mathrm{Dir}\left(\alpha\right) = f\left(\alpha^k, y^k\right) = \frac{\Gamma\left(\alpha_0\right)}{\prod_{i=1}^{k}\Gamma\left(\alpha_i\right)}\prod_{i=1}^{k} y_i^{\alpha_i - 1} \tag{3.2}$$

where α^k is the parameter vector of positive reals containing k components, y_i is the associated vector element, and α_0 is the summation $\alpha_0 = \sum_{i=1}^{k}\alpha_i$ [Lin, 2016]. SparCC thus samples from this distribution function to estimate the true fractions from the obseved compositional counts.

If the set of data is considered as the vectors for the components i (in this case each OTU is a component), then the fractions distribution is the Dirichlet distribution [Gelman et al., 2013]:

$$p\left(\underline{x}|\underline{N}\right) = \mathrm{Dir}\left(\underline{N} + 1\right), \tag{3.3}$$

where $\underline{x}$ is the components' true fractions and $\underline{N}$ is the observed counts. Note the $\underline{N} + 1$ component in the Dirichlet distribution; this added value is defined as adding a *pseudocount* of 1 to all count values for each component[1].

3.1.2 Log Ratio

Compositional data is frequently normalized by employing the use of the centered log-ratio. Consider an $m \times n$ matrix where the m rows are samples, and the n columns are features of the data. Since this theory is applied to 16S rRNA compositional data, the theory will be presented in relation to 16S data, but any type of compositional data should be applicable. Thus, imagine an OTU table with m samples, and n OTU IDs.

The log-ratio transformation is defined as:

$$y_{ij} = \log\frac{x_i}{x_j} = \log x_i - \log x_j, \tag{3.4}$$

where y_{ij} contains the new information of the true abundances of OTUs since the fraction of OTU x_i and OTU x_j is equal to the ratio of the true abundances. Additionally, the log-ratio transformation allows the resulting value to be independent of the other OTUs in the analysis, and y_{ij} is no longer limited to the simplex of the dimensionality of the components i and j in the range $\{1, 2, ..., n\}$.

3.1.3 Compositional Based Dependencies

From here, the value that should be used in the analysis comes from the variance of the log-ratio applied across all samples [Aitchison, 2003]:

[1] Pseudocounts are used in compositional data to avoid division by zero errors that could occur with generally sparse data. They are usually small values compared to the rest of the data in order to minimize the effects of influencing fraction calculations. Different methods have suggested various optimal pseudocount values, but the authors use a pseudocount of 1 for this method. For more examples of pseudocounts refer to their usage in these papers: Weiss et al. [2017], Mandal et al. [2015], Wang et al. [2017].

$$t_{ij} \equiv \mathrm{Var}\left[\log \frac{x_i}{x_j}\right] = \mathrm{Var}\left[y_{ij}\right]. \tag{3.5}$$

Aitchison uses this value t_{ij} to describe the dependencies between the components and calls this resulting matrix the *variation matrix* ($\mathbf{T}$). Note that this relation will result in $t_{ij} = 0$ for perfectly correlated OTUs and will vary with other correlation relations [Friedman and Alm, 2012]. A basic property of variance and covariance for the sum of two random variables X and Y leads to the relation:

$$\mathrm{Var}\left[aX - bY\right] = a^2\mathrm{Var}\left[X\right] + b^2\mathrm{Var}\left[Y\right] - 2ab\mathrm{Cov}\left[X, Y\right], \tag{3.6}$$

where a and b are both constants. This property will be applied in Equation 3.7.

3.1.4 Basis Correlation Relation

In order to better understand the scaling of the dependency value and determine what constitutes a weak or strong dependence, we need to relate the value to the correlation of the true abundances in the compositional data. To relate Equation 3.5 to the basis correlation, first apply the relation of Equation 3.6:

$$\begin{aligned} t_{ij} \equiv \mathrm{Var}\left[\log \frac{x_i}{x_j}\right] = \mathrm{Var}\left[\log \frac{w_i}{w_j}\right] = \mathrm{Var}\left[\log w_i - \log w_j\right], \\ = \mathrm{Var}\left[\log w_i\right] + \mathrm{Var}\left[\log w_j\right] - 2\mathrm{Cov}\left[\log w_i, \log w_j\right], \end{aligned} \tag{3.7}$$

where $x_i = w_i$, $x_j = w_j$, and $a, b = 1$. Then we can obtain the equation:

$$t_{ij} = \omega_i^2 + \omega_j^2 - 2\rho_{ij}\omega_i\omega_j, \tag{3.8}$$

where ω_i^2 and ω_j^2 are the variances of the log-transformed variances of components i and j ($\mathrm{Var}\left[\log w_i\right]$ and $\mathrm{Var}\left[\log w_j\right]$ respectively), and ρ_{ij} is the correlation between them.

Friedman and Alm's aim is to infer the unobserved covariance matrix from the variation matrix T, but this is not possible in the general case because the basis variances are unknown. So an exact solution is not possible, but an approximation can be found[2]. We initially rewrite Eq. 3.8 for ρ_{ij}:

$$\rho_{ij} = \frac{\omega_i^2 + \omega_j^2 - t_{ij}}{2\omega_i\omega_j}, \tag{3.9}$$

which gives the basis correlations if we have the variances ω_i and ω_j of the log-transformed basis and the quantity t_{ij}. To solve for these variables start with finding

[2]The approximation is valid for situations when there are many components that are sparsely correlated. This is assumed for sparse compositional data as seen in Eq. 3.13. If we were to assume that all of the basis variables have the same variance ω, then we would have the relation $2 \gg 2\langle\rho_{ij}\rangle_i$ and the **average** correlations would be small versus needing a specific **set of correlations** to be small [Friedman and Alm, 2012].

the approximation for the basis variance by defining the variation of component i as:

$$t_i \equiv \sum_{j=1}^{D} t_{ij} = d\omega_i^2 + \sum_{j \neq i} \omega_j^2 - 2 \sum_{j \neq i} \rho_{ij} \omega_i \omega_j, \tag{3.10}$$

across all variances D where $d \equiv D - 1$. Then we factor Eq. 3.10 by the term $d\omega_i^2$:

$$t_i = d\omega_i^2 \left[1 + \frac{1}{d} \sum_{j \neq i} \frac{\omega_j^2}{\omega_i^2} - 2 \frac{1}{d} \sum_{j \neq i} \rho_{ij} \frac{\omega_j}{\omega_i} \right], \tag{3.11}$$

and then use $\langle \cdot \rangle_i$ notation to stand for the averaging of all pairs corresponding to i. Thus t_i is now:

$$t_i \equiv d\omega_i^2 \left[1 + \langle \left(\frac{\omega_j}{\omega_i} \right)^2 \rangle - 2 \langle \rho_{ij} \frac{\omega_j}{\omega_i} \rangle_i \right], \tag{3.12}$$

and because we make the assumption that the correlation terms ρ_{ij} are small:

$$1 + \langle \left(\frac{\omega_j}{\omega_i} \right)^2 \rangle_i \gg 2 \langle \rho_{ij} \frac{\omega_j}{\omega_i} \rangle_i, \tag{3.13}$$

we are left with an expression for t_i containing only the variances of the log-transformed basis:

$$t_i \simeq d\omega_i^2 + \sum_{j \neq i} \omega_j^2 \,, \; i = 1, 2, ...D. \tag{3.14}$$

3.1.5 Statistical Significance

In FastSpar, statistical significance is calculated by sampling permutations of the data with replacement [Watts et al., 2018][3]. The p-value implemented is termed the *exact p-value* p_e [Phipson and Smyth, 2010]. It's derived from a random sample of m permutations with replacement, and the test statistics t_{perm} may contain both repeat values and original observed values t_{obs}. Note that p_e will be slightly less than $\frac{b+1}{m+1}$ for all $b = 1, 2, ..., m$ because the original data may be present at least once in one of the permutations.

If we let the exact permutation value p_e be:

$$p_e = P\left(B \leq b\right), \tag{3.15}$$

where B is the number of permutations out of m that have test statistics at least as extreme as t_{obs}, then we can find p_e for any given b. Now let B_t be the number of unique statistic values greater than t_{obs}, and consequently $p_t = \frac{B_t+1}{m_t+1}$. Assuming a true null hypothesis H_0, and $B_t = b_t$, then B_t will follow a discrete uniform

[3]This is described briefly in the paper, but can be found via the author's GitHub repository home directory *README.md* file and source code in the *pvalue.cpp* script. Refer to the GitHub repository here: `https://github.com/scwatts/fastspar`.

distribution over the integers $0, ..., m_t$ and B follows a binomial distribution with size m and probability p_t. This is formally defined as:

$$p_e = \sum_{b_t=0}^{m_t} P\left(B \leq b | B_t = b_t\right) P\left(B_t = b_t | H_0\right) = \frac{1}{m_t + 1} \sum_{b_t=0}^{m_t} F\left(b; m, p_t\right), \tag{3.16}$$

where $F()$ is the cumulative probability function of the binomial distribution [Phipson and Smyth, 2010].

3.1.6 SparCC and FastSpar

By following the previous steps for the t_i approximation, the basis correlations ρ_{ij} can be calculated by using Eq. 3.9. The authors of SparCC initially suggest a basic procedure to infer the correlations which follows as such:

1. Estimate the component fractions with the Dirichlet distribution to obtain the fractions matrix X
2. Compute the variation matrix T and the variations t_i
3. Plug the log-basis variances ω_i from Eq. 3.14 into Eqs. 3.7 and 3.8 to obtain ρ_{ij}
4. Iterate through steps 1-3, while each time identifying the most strongly correlated pairs of components and discarding those that form exclusive pairs
5. Estimate the fractions of remaining components and compute the new t_i, ω_i, and ρ_{ij}

The FastSpar implementation, which this thesis utilizes, follows the above procedure but differs when assessing statistical significance as outlined in Section 3.1.5. Thus, following the above-mentioned procedures we obtain correlation values and their respective p-values. With this information in hand, we may analyze the newly generated correlation networks.

3.2 NetShift

The NetShift methodology presents a graph theoretic approach to quantifying differences in correlation networks. With its standard approach, it is possible to compare two similar types of networks.

3.2.1 General Properties

Given a graph $\boldsymbol{G}(\mathcal{V}, E)$, consider the general properties: degree, density, average path length, and cluster coefficient. The degree of a node i in a graph is equivalent to the number of edges connected to the node and denoted by k_i. In a graph with nodes that are not self-connected, the maximum number of possible edges E is limited to

the number of nodes in the network as depicted in the relation $|\mathcal{V}|\left(|\mathcal{V}|-1\right)$. The graph density is defined as:

$$D=\frac{2|E|}{|\mathcal{V}|\left(|\mathcal{V}|-1\right)}, \tag{3.17}$$

where $\mathcal{V}$ and E are the number of nodes and edges respectively. D is dependent upon how connected the network is, and if the network is fully connected, $D=1$. Now if we consider a node in the network v_1 and another node v_2, we can find the shortest path that connects the two. If the two are directly connected then $d\left(v_1,v_2\right)=1$, if there are multiple nodes between the two $d\left(v_1,v_2\right)$ will be the number of connecting edges, and if there is no path between the two $d\left(v_1,v_2\right)=0$. Thus we can then define the average path length l_G of the network as:

$$l_G=\frac{1}{n\left(n-1\right)}\sum_{i\neq j}d\left(v_1,v_2\right). \tag{3.18}$$

Next, we will define the following properties: clustering coefficient, betweenness centrality, and coreness centrality. The local clustering coefficient is defined for each individual node in the network. It is dependent upon the number of edges connecting a node to its neighbors compared to the possible number of edges that it could have. For each node i, there exists a neighborhood N_i which are the node's neighbors:

$$N_i=\{v_j : e_{ij}\in E\vee e_{ji}\in E\}. \tag{3.19}$$

Using the neighborhood, the local clustering coefficient is defined as:

$$C_i=\frac{2L_i}{k_i\left(k_i-1\right)}, \tag{3.20}$$

where L_i is the number of edges in the neighborhood N_i of node i [Watts and Strogatz, 1998, Barabási and Pósfai, 2016]. Additionally, the average of the network clustering coefficients $\overline{C}$ can be calculated by averaging over all of the clustering coefficients:

$$\overline{C}=\frac{1}{n}\sum_{i=1}^{n}C_i. \tag{3.21}$$

Centrality measures in graphs aim to assign a value of importance in the network. One such relevant metric for real systems is the betweenness centrality, which is related to the number of shortest paths l_G that pass through a given node. The higher the betweenness centrality, the more the node interacts with the rest of the network. This is an important metric for real-life systems as these nodes may drive the behavior of the network. Betweenness centrality is defined as:

$$g\left(\mathcal{V}\right)=\sum_{i\neq\mathcal{V}\neq j}\frac{\sigma_{ij}\left(\mathcal{V}\right)}{\sigma_{ij}}, \tag{3.22}$$

where σ_{ij} is the number of shortest paths from i to j, and $\sigma_{ij}\left(\mathcal{V}_i\right)$ is the number that pass through $\mathcal{V}$ [Freeman, 1977]. Similarly to Equation 3.21, we let the average betweenness centrality $\overline{g(\mathcal{V})}$ be:

$$\overline{g(\mathcal{V})}=\frac{1}{n}\sum_{k=1}^{n}g\left(\mathcal{V}\right)_k. \tag{3.23}$$

Another relevant metric is coreness centrality which observes the k-shell indices of a node's neighborhood [Bae and Kim, 2014]. By taking a graph and its nodes, the k-shell indices are assigned by first removing all nodes from the graph whose degree is equal to 1. These are then assigned to the 1-core with the index 1. This method is done recursively, increasing k each time until all nodes have been assigned to a k-core with a respective k-shell index between 1 and k [Bae and Kim, 2014, Lü et al., 2016, Dorogovtsev et al., 2006]. After assigning the k-shell indices, each node $\mathcal{V}$ can be assigned a neighborhood coreness value C_{nc}:

$$C_{nc}\left(\mathcal{V}\right) = \sum_{w \in N(\mathcal{V})} \text{k-s}\left(w\right), \tag{3.24}$$

where w is in the neighborhood of $\mathcal{V}$, and k-s(w) is the k-shell index of node w. Either C_{nc} or the extension of C_{nc} to include the neighbors of the neighbors of $\mathcal{V}$ can be used as a metric.

3.2.2 Quantifying Community Structure Change

While we could compare the two networks and their respective properties denoted in Section 3.2.1, these cannot be used solely to compare two networks if they contain different nodes. So, as stated previously, the network comparison needs to address the common sub-graphs of the networks.

By comparing the two sets, a baseline score called the Jaccard Edge Index (JEI) can be found to quantify the re-wiring of edges, and highlight major shifts in the role of nodes in the networks. JEI is defined as:

$$\text{JEI} = \frac{A_E \cap B_E}{A_E \cup B_E}, \tag{3.25}$$

where A_E and B_E are the edge sets for a control network A and case network B respectively. The authors suggest that the Jaccard index can be useful, but it does not describe the change between network A and B sufficiently so that the directionality of the changes is taken into account [Kuntal et al., 2018]. For example, there may be a shift in the edges between a node in A, and the same node in B. Consider the case where we are observing the same node n in the two networks (A^n and B^n). Then the edge list of A^n_E may not be equivalent to the edge list of B^n_E. The authors thus propose a new score called the Neighbor shift (NESH) which looks at the individual nodes in networks A and B, and assigns them a score based upon the change of the node's neighborhood in the two networks.

The author's introduce a few components to be included in the NESH calculation. They define NESH as:

$$\text{NESH} = 1 - \left(X - \left(Y + Z\right)\right), \tag{3.26}$$

where X is a measure similar to the JEI but focusing on the node's neighbors, and Y, Z are penalty terms for X that consider exclusive enrichment of the case set B over the control A.

If we let $[Neighbors]^A$ and $[Neighbors]^B$ be the set of first neighbors for a shared node in A and B respectively, then we define the intersection of the node for set A

and B with $[Neighbors]^A \cap [Neighbors]^B$; the union as $[Neighbors]^A \cup [Neighbors]^B$; the relative complement of set B (objects in B but not in A) as $[Neighbors]^B - [Neighbors]^A$; and the maximum degree of the nodes in B as $\max\left(k^B\right)$. Accordingly, X, Y, and Z are defined as:

$$\begin{aligned} X &= \frac{[Neighbors]^A \cap [Neighbors]^B}{[Neighbors]^A \cup [Neighbors]^B}, \\ Y &= \frac{[Neighbors]^B - [Neighbors]^A}{\max\left(k^B\right)}, \\ Z &= \frac{[Neighbors]^B - [Neighbors]^A}{[Neighbors]^B \cup [Neighbors]^A}. \end{aligned} \tag{3.27}$$

Specifically, Y represents the unique connections in B compared to the possible number of connections that the node can have. Z defines the quantity the exclusive set over the union of the interacting partners. NESH is then calculated by substituting Equations 3.27 into Equation 3.26.

4

Methods

This chapter discusses the methods surrounding the research from this thesis. Section 4.1 describes the data acquisition and pre-processing steps for the meta-analysis data. Section 4.2 presents the generation of the correlation values and their respective statistical significance values. Section 4.4 discusses the network analysis pipeline. Section 4.5 explains the visualization methods that were implemented to allow for visual and quantitative comparison of the networks and their structure. Unless otherwise noted, analysis work was performed on a remote Amazon Web Service Elastic Compute Cloud instance running RStudio RServer and Python 3. Network visualization was performed in Cytoscape [Shannon et al., 2003].

4.1 Data Acquisition and Pre-processing

This analysis uses 28 publicly available 16S rRNA human gut microbiome studies that were gathered and curated by Duvallet et al. [2017] and available with some minor changes to the open-source code made available in her GitHub repository [Duvallet, 2018]. The respective studies are listed in Table 4.1 along with their associated case disease, and the number of unique samples for the control and case cohorts. As mentioned previously, sequenced 16S rRNA data can come in various formats depending upon the methods that individual labs use. It is important to consider that community structure from interpreted results differs based upon the V-region targeted in sequencing [Teng et al., 2018]. So to avoid study-based artifacts, Duvallet et al. collapsed the resulting OTU's to the genus level.

4.1.1 Acquisition

The raw study data can be acquired through communication with the various authors or via the National Center for Biotechnology Information (NCBI) and European Nucleotide Archive (ENA) databases and the respective accession number for the study. The data used in this thesis came directly from Duvallet et al.'s processing pipeline. We used their data because their study indicated that there was a clear signal between healthy and diseased microbiomes. With this knowledge we assumed that there would be some type of community structure difference between the healthy and diseased correlation networks.

Dataset ID	Control	Controls (N)	Case	Cases (N)
Singh et al. [2015]	H	82	EDD	201
Schubert et al. [2014]	H	154	CDI	93
Schubert et al. [2014]	H	154	nonCDI	89
Vincent et al. [2013]	H	25	CDI	25
Youngster et al. [2014]	H	4	CDI	19
Goodrich et al. [2014]	H	428	OB	185
Turnbaugh et al. [2008]	H	61	OB	195
Zupancic et al. [2012]	H	96	OB	101
Ross et al. [2015]	H	26	OB	37
Zhu et al. [2013]	H	16	OB	25
Baxter et al. [2016]	H	172	CRC	120
Zeller et al. [2014]	H	75	CRC	41
Wang et al. [2011]	H	54	CRC	44
Chen et al. [2012]	H	22	CRC	21
Gevers et al. [2014]	nonIBD	16	CD	146
Morgan et al. [2012]	H	18	UC, CD	108
Papa et al. [2012]	nonIBD	24	UC, CD	66
Willing et al. [2010]	H	35	UC, CD	45
Noguera-Julian et al. [2016]	H	34	HIV	205
Dinh et al. [2014]	H	15	HIV	21
Lozupone et al. [2013]	H	13	HIV	23
Son et al. [2015]	H	44	ASD	59
Kang et al. [2013]	H	20	ASD	19
Alkanani et al. [2015]	H	55	T1D	57
Mejía-León et al. [2014]	H	8	T1D	21
Wong et al. [2013]	H	22	NASH	16
Zhu et al. [2013]	H	16	NASH	22
Scher et al. [2013]	H	28	PSA, RA	86
Zhang et al. [2013]	H	25	CIRR, MHE	46
Scheperjans et al. [2014]	H	74	PAR	74
Total:		1816		2210

Table 4.1: Table containing information on the respective studies used in this thesis. Also listed are the diseases investigated and the respective control and case cohort sizes. The respective acronyms are defined as: colorectal cancer (CRC), non-alcoholic steatohepatitis (NASH), Parkinson's disease (PAR), ulcerative colitis (UC), *Clostridium difficile* infection (CDI), psoriatic arthritis (PSA), liver cirrhosis (CIRR), healthy (H), enteric diarrheal disease (EDD), Crohn's disease (CD), autism spectrum disorder (ASD), type I diabetes (T1D), liver disease (LIV), human immunodeficiency virus (HIV), inflammatory bowel disease (IBD), arthritis (ART), obesity (OB), minimal hepatic encephalopathy (MHE), rheumatoid arthritis (RA), control patients with diarrhea who tested negative for CDI (nonCDI), and control patients with gastrointestinal symptoms but no intestinal inflammation (nonIBD).

Duvallet et al. collected the raw *FASTA* and *FASTQ* files for the studies. This was run through a standardized platform to remove barcodes, and primers, and handle multiplexed files accordingly[1]. This pipeline uses clustering at 100% similarity with USEARCH, and the naive Bayes RDP classifier to assign taxonomy [Wang et al., 2007]. The resulting processed raw data has been posted on Zenodo[2]. At this point, the data is now in standard OTU table format where it contains the counts associated for each genus in each sample. The data was then extracted from the beginning of Duvallet et al.'s MicrobiomeHD pipeline during the concatenation of all of the study data. To extract the data we ran the beginning of the MicrobiomeHD pipeline, but left out the sample normalization step to retain whole counts in the resulting concatenated dataset. We included the author's collapsing to the genus-level and automatic discarding of unassigned taxa[3]. At this point, we wrote a file containing all OTU data, and a file containing the concatenated table of all metadata (which includes various information such as class information, sample ID's, NCBI and ENA accession numbers, and other information).

4.1.2 Pre-processing

Referencing Table 4.1, the total number of samples in the data set is $n_{total} = 4026$. Of these, there are $n_{control} = 1816$ control samples and $n_{case} = 2210$ case samples. At this stage, some of the samples were excluded due to insufficient healthy and diseased requirements outlined by Duvallet et al.. After excluding the samples that should not be considered in the healthy versus diseased sets, we are left with $n_{control} = 1751$ and $n_{case} = 1973$. In the OTU data, there are 291 features (genera) containing the counts for the respective taxa in each sample. Some of the analyses and techniques implemented in this study use the data as-is, in its raw count form, and some require additional processing.

Prior to the FastSpar and NetShift analysis we split the data into the healthy and diseased sets and then filter the OTUs so that each one is present in at least 5% of the samples in the respective data set. We employ this method to avoid possible division by 0 errors, and to use taxa that are actually present across a large majority of the samples. It also eliminated a potential way for our statistical significance calculations to be biased. In most cases these taxa have either no or a very low variance in their counts, and keeping taxa with less than a 5% abundance often time limits the number of unique permutations to 1. Essentially, this is another step

[1]Information on the pipeline is available here: `https://amplicon-sequencing-pipeline.readthedocs.io/en/latest/index.html`.

[2]This is linked to in Duvallet [2018], but can be directly accessed here: `https://zenodo.org/record/1146764#.XNBJ1o5KhPY`.

[3]We noticed that the unassigned data represented between 0-50% of the data for a given sample. This represents a significant portion of data and since the taxa annotations were based upon an older database, some portion of this unassigned data could represent newly discovered or undiscovered microbes. The other portion of the data is probably from corrupt or In an attempt to answer this, we tried to run all of the raw studies through our in-house pipeline. Unfortunately we were not able to finish this effort due to a time constraint limiting our ability to wrangle all of the raw 16s data from the various studies.

to ensure artifacts do not influence the end result of the analyses. In total, there were 107 and 111 genera after filtering the features to meet the 5% threshold in the control (healthy) and case (diseased) set respectively.

4.2 Correlation Estimation

Due to SparCC's unfavorable runtime, memory usage, and high false-positive rate, we used the FastSpar C++ command line interface (CLI) implementation of SparCC with improved runtime, memory usage, and statistical significance methods [Friedman and Alm, 2012, Watts et al., 2018]. FastSpar is run using the raw count data which must be converted to BIOM tsv format[4] and requires the headers to contain the sample ID and the row names to be the OTUs. We run the toolkit on our 5% occurrence threshold data for the combined case and control set, and the respective separated case and control sets.

After following the initial correlation process, we calculate the exact p-value. First, 10000 bootstrap samples were generated and then we inferred correlations for each of the bootstraps in parallel. The resulting correlations were then used to calculate the exact p-value for each correlation in the original calculation.

4.3 Correlation and Extended Correlation Filtering

Per the arbitrary threshold selection method in the literature, we specifically follow Friedman and Alm [2012]'s method to make the matrix sparse by employing value thresholding to reduce the number of noisy correlations. In this step we remove the self-association connections to create less dense networks with large correlation magnitudes.

To include the statistical significance metric generated by FastSpar in this thresholding step we ran the algorithm and performed 10000 and 100000 bootstrap permutations in order to calculate the exact p-values. We first ran 10000 bootstrap permutations, and saw that the significant correlation values were the same across the samples with a majority of the p-values decreasing from 0.0001 to 0.00001 when we ran the 100000 bootstrap permutations. Initially, we thought that by increasing bootstraps we would be able to weed out additional connections. It appears that a majority of the correlations in our data are statistically significant even with the false discovery rate correction employed by FastSpar.

We then investigated filtering the data based upon statistical significance and correlation value. Our analysis included a combination of threshold measures: using the fully connected matrices (no filtering of correlation value or p-value, keeping all

[4]The description of the BIOM format can be found here: `http://biom-format.org/documentation/biom_format.html`

connections with a correlation value threshold of $|c| \geq 0.25$; keeping all connections with a p-value threshold of $p \leq 0.01$ and correlation value c of $|c| \geq 0.25$; and keeping all connections with a p-value threshold of $p \leq 0.01$ and correlation value of $|c| \geq 0.05$. In the networks, if there were nodes that were no longer connected, they were removed for the downstream analysis.

After applying these filtering methods, we became aware that the arbitrary threshold selection might leave out comparable connections shared between the networks. For example, let there be a connection between genera $A \Leftrightarrow B$ in the healthy network with a correlation of 0.26 as demonstrated in Figure 4.1. It will be selected when we filter according to our method. Now let the same connection exist in the diseased network, but with a correlation value of 0.24. The connection in the diseased network will be dropped despite the difference between the two being relatively small. This may be eliminating an important feature in the overlapping sub-graph

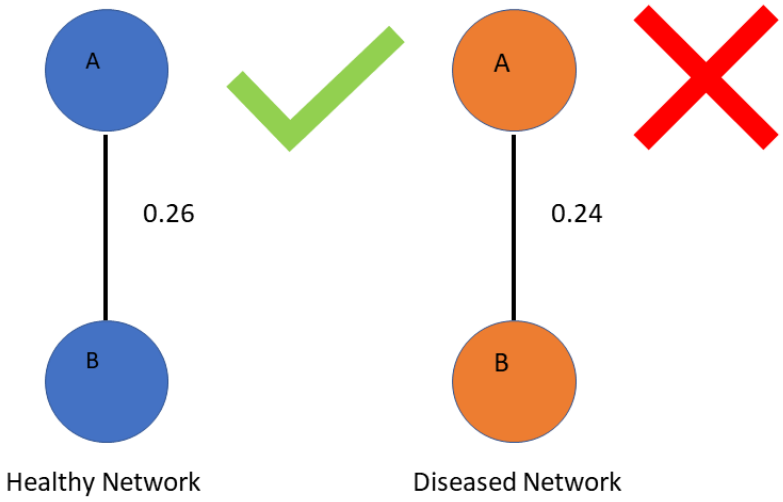


Figure 4.1: Example of potential artifact from single arbitrary threshold. Given an example edge in the healthy network that meets the correlation threshold and the same edge in the diseased that just barely misses it, the single threshold value would eliminate this edge from further analysis.

of the two networks. To remedy this potential artifact generation, we decided to check connection overlaps between the diseased and healthy networks and select an extended correlation value b. We first filtered one of the networks by our filtering methods, and then checked all shared connections with the unfiltered network. We took the correlation value c_f that we filtered the network by and checked to see if the unfiltered network shared connection correlation was above the new correlation filter value c_{fnew} defined as $c_{fnew} = c_f - b$. In the case that the unfiltered network connection correlation value is above c_{fnew}, we appended it to the unfiltered network's filtered results. We repeated the process for the other network so that, after normal filtering and the second filtering were complete, we had two filtered networks that contained the added connections that met the second correlation filter criteria.

After trying many values we implemented a second correlation selection window that was set to 0.1. The edges that met the extended threshold were included in the respective networks and the resulting networks were used in the downstream analysis.

4.4 Network Analysis

After converting the correlation and statistical significance matrices to edge lists, we computed the standard graph topology scores. This was performed for the combined case and control data, as well as the separate control and case networks. Then we utilized the NetShift web tool from Kuntal et al. [2018] and applied it to the control and case networks. Since NESH is direction-agnostic, and only a network topology feature, we do not include correlation values in the calculation.

4.5 Visualization

To visualize our data we used the open source platform Cytoscape [Shannon et al., 2003]. When writing the graph files, we assigned Hex color codes to the unique genera and families respectively, to aid in differentiating nodes in the network. Family and genera colors were kept the same across the full data set and the control and case sets. In Cytoscape, we set the negative correlation values to blue, and positive to red while scaling the size of the edge so that the larger the absolute value of the correlation $|c|$ is the larger the edge is.

We include our own visualization of the networks with our correlation matrices, but we also used the visualizations obtained from the web application provided by Kuntal et al.[5] for visualizing the network shift. Community shuffle plots from the application are also presented as a tool to better understand the graph visualization of the network shift. The network topology scores and features are listed in tables in the results, and the importance of the metrics will be further discussed there.

[5]The web application can be found here: `https://web.rniapps.net/netshift/`

5

Results

This chapter presents the results of the methods described in Chapter 4. We begin with Section 5.1 which presents the filtering method results and the scope of the network that will be further analyzed. Section 5.2 presents the extended filter and the further reduction of the networks.

5.1 Initial Correlation and Statistical Significance Filtering

Using the FastSpar method, we generated fully connected correlation matrices for the control and case data. A heat map visualization of the healthy network can be found in Figure 5.1 and the diseased heat map can be found in the appendix as Figure A.1. The genera in the network are listed on the respective axes, and spaces are colored according to the correlation values present in the matrix. Self-associated connections occur with a correlation strength of 1; as is visible along the diagonal of the heat map. These figures show fully connected correlation networks, so it is necessary for us to filter this initial result for better understanding of the underlying network properties. Figures 5.2 and 5.3 present the correlation and p-value histograms for the healthy (control) and disease (case) networks with some of the filtering criteria utilized.

In Figure 5.2, a majority of the correlation values are significant. In the unfiltered data, the distribution of correlations appears to be close to a normal distribution, with some skew toward 1.0. Almost all of the p-values have a value of 0.00001. 107 of these p-values were equal to 1, and they were associated with a taxon's self correlation (as mentioned previously with regard to Figure 5.1). We were not interested in correlations close to 0, nor were we interested in the correlation of a taxa with itself, so in the next row of Figure 5.2 we filtered the data by keeping correlations $|c| \geq 0.25$, but did not filter by the p-value. After this step we saw a large reduction in the number of correlation values, as well as p-values that were between our smallest and largest values. Filtering at this level consequently has revealed that roughly a quarter of the remaining correlations are associated with p-values of 1. We knew that these belong to the self-correlations of the taxa, so we changed our filtering criteria to keep the same correlation filter while adding a filter to keep all p-values $p \leq 0.01$ and eliminating the diagonal (of the matrix) self-

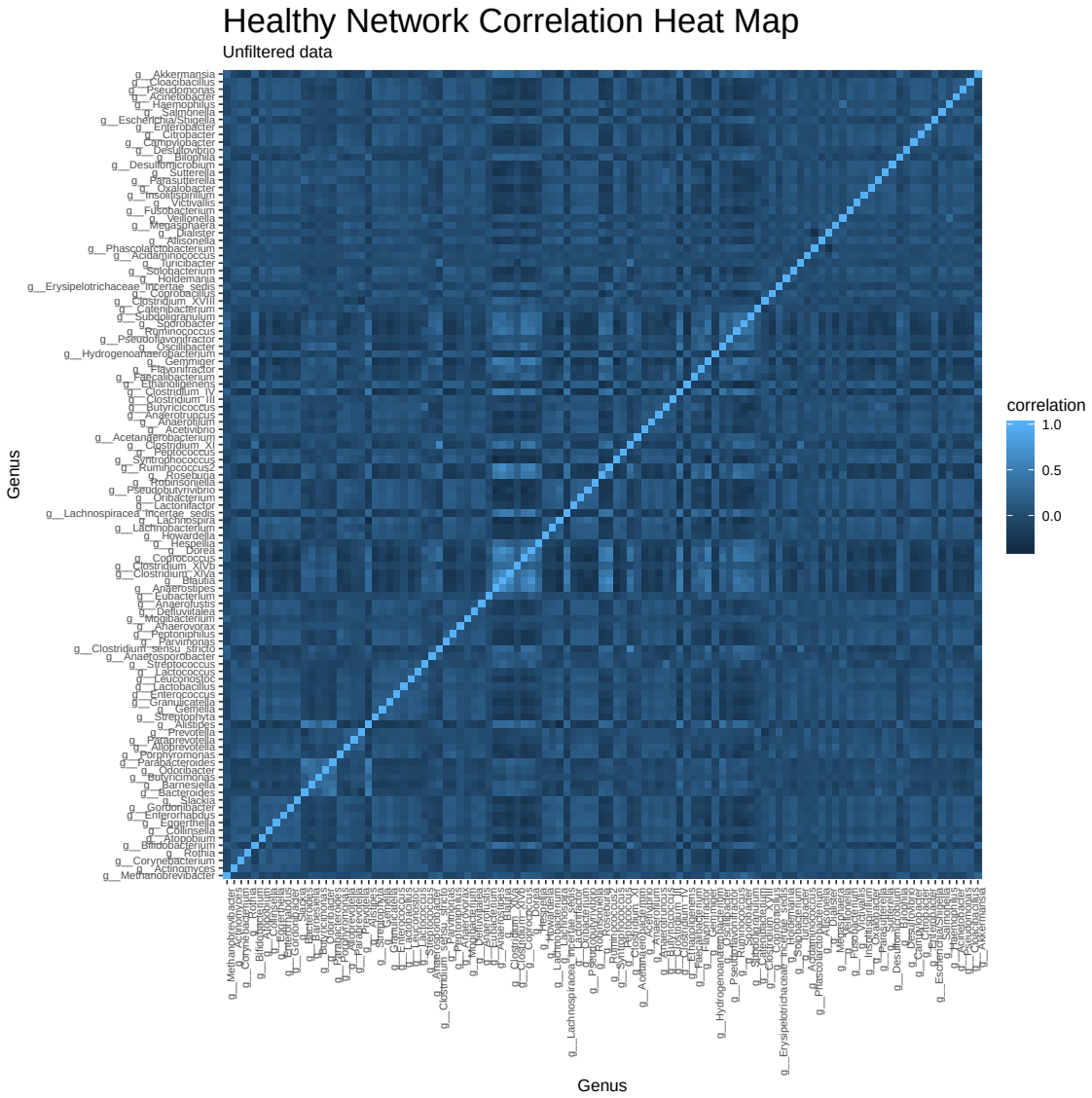


Figure 5.1: Heat map of the resulting FastSpar correlation matrix. Genera are listed on the respective axes and correlation values are colored based upon their strength.

associated correlations. At this step we significantly reduced the connectivity and size of the network by removing correlation values close to 0 while removing all of the statistically insignificant correlations. We ran many different filtering methods, but this one appeared to be the closest to what we were looking for.

Included in the bottom row of Figure 5.2 is an additional filtering method where we filtered by $p \leq 0.01$ and $c \geq 0.05$. With this filtering method there are still several thousand connections in the network, but all the p-values are statistically significant. Since there were still so many correlation values, we must increase the correlation threshold.

The distributions listed in Figure 5.3 are similar to those of the healthy network. We applied the same filtering methods as described before due to the similarity in

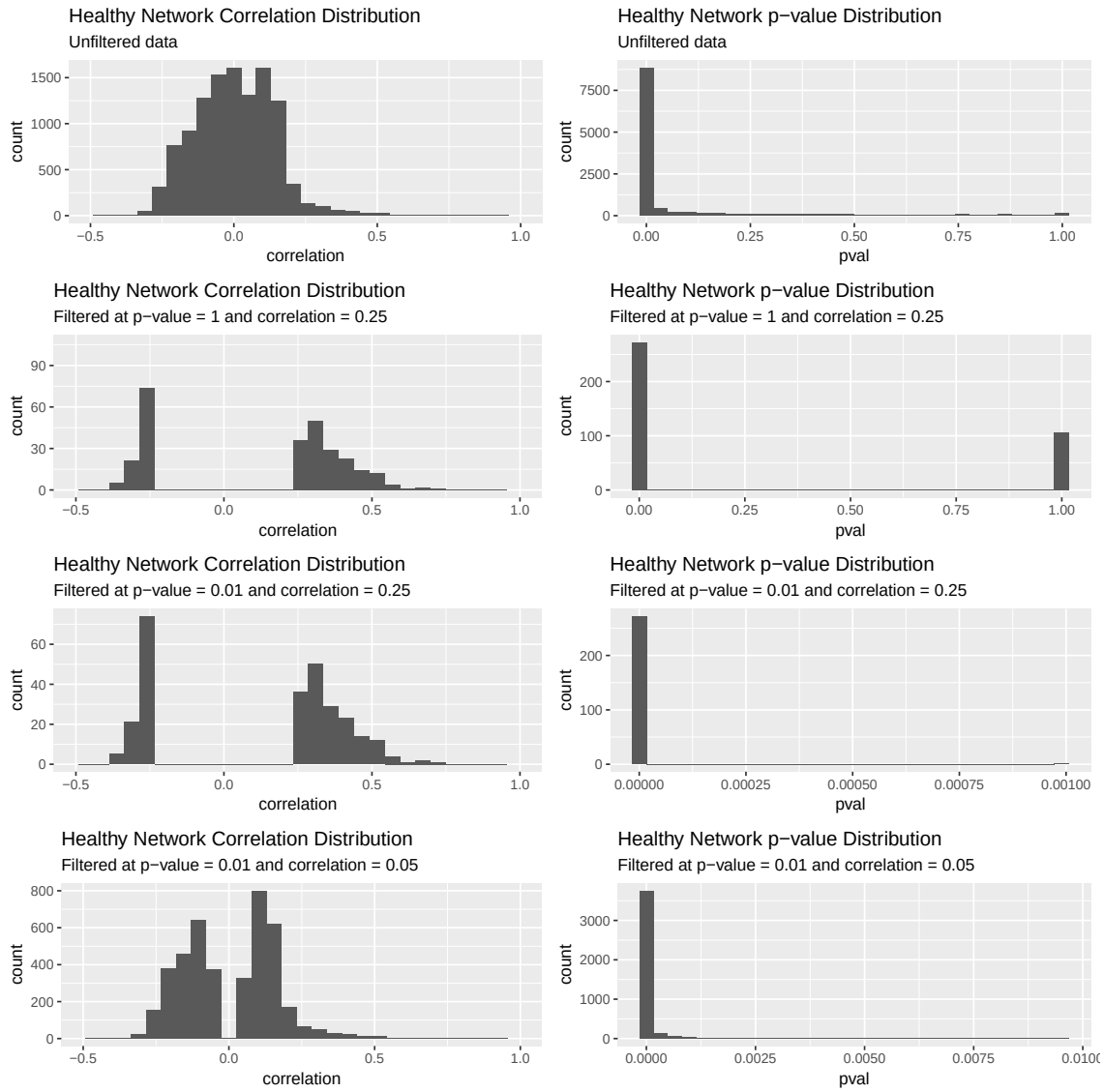


Figure 5.2: Histogram distribution comparison of the healthy network filtered at varying correlation and p-value levels. The chosen parameters selected for visualization include: the raw unfiltered data, the network containing all correlations c where $|c| \geq 0.25$, network with all connections with $p \leq 0.01$ and $|c| \geq 0.25$, and the network with all connections with $p \leq 0.01$ and $|c| \geq 0.05$.

correlation and p-value distributions. In the diseased network there were 111 genera present, and thus 111 self-associated connections were filtered out. We found that by choosing the p-value cutoff of $p \leq 0.05$ and the correlation cutoff of $|c| \geq 0.25$ we were able to eliminate a majority of the connections and the remaining connections were all statistically significant. Similar behavior between the networks in this filtering stage suggests that the structure of the two networks is similar and that further comparisons might yield additional results.

Table 5.1 describes the number of edges before and after applying the initial filter. There were a total of 5671 and 6105 unique edges for the healthy and diseased

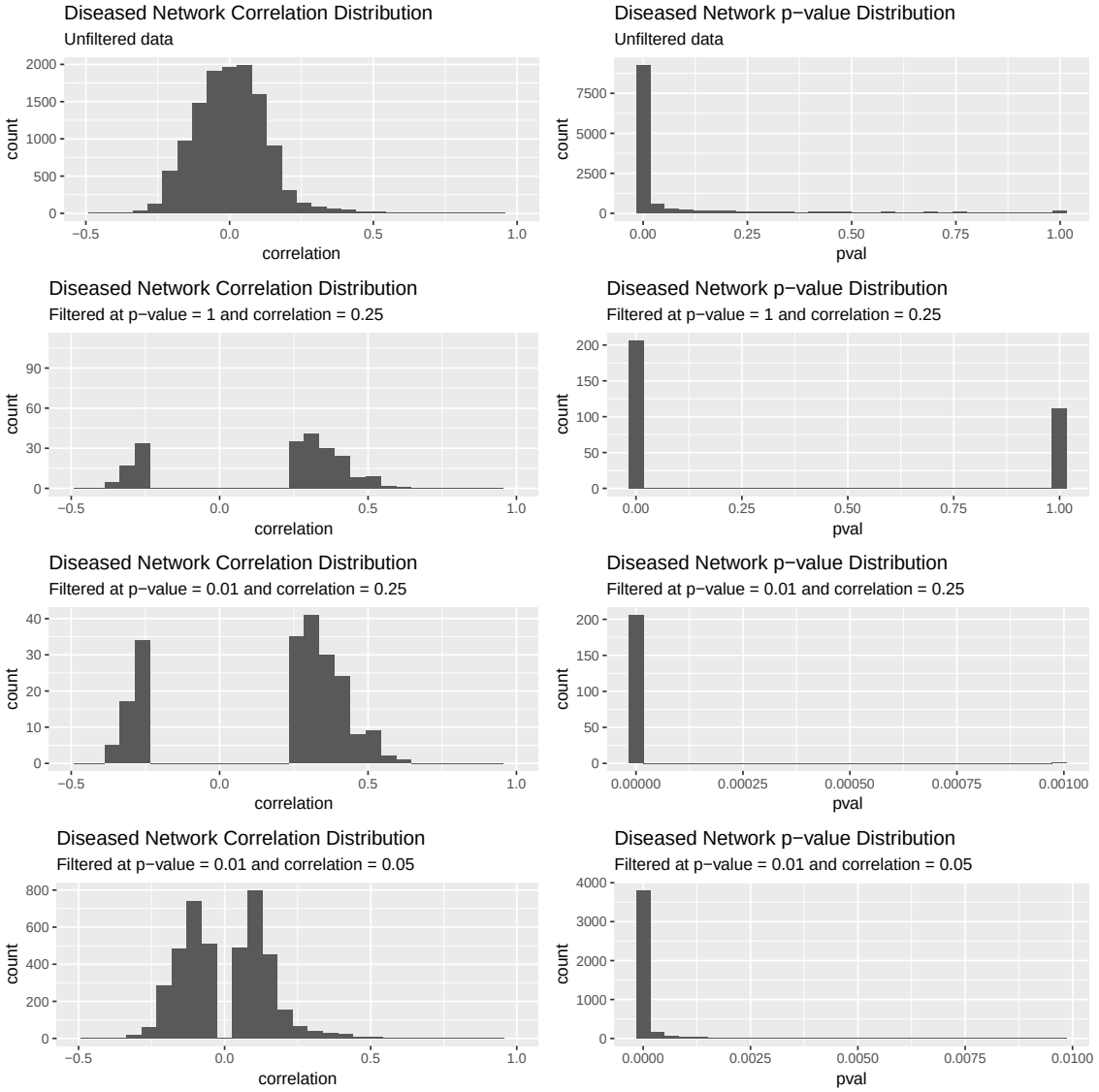


Figure 5.3: Histogram distribution comparison of the diseased network filtered at varying correlation and p-value levels. The chosen parameters selected for visualization include: the raw unfiltered data, the network containing all correlations c where $|c| \geq 0.25$, network with all connections with $p \leq 0.01$ and $|c| \geq 0.25$, and the network with all connections with $p \leq 0.01$ and $|c| \geq 0.05$.

networks respectively. After the first round of filtering there were 272 and 206 edges respectively. Despite there being more connections in the diseased network prior to filtering, the first filtering step yielded more healthy edges.

5.2 Extended Correlation Filtering

In Section 4.3 we mentioned the implementation of an extended correlation filter for the correlation values close to the original filter's threshold. We found the shared

Cohort Data	Total Edges	Edges After Filter 1
Healthy	5671	272
Diseased	6105	206

Table 5.1: Table including the total edges before and after the first filter. We first found the total edges in each cohort by only looking at the upper triangle of the respective matrix. We then filtered with the 0.25 correlation threshold and the p-value threshold of 0.01.

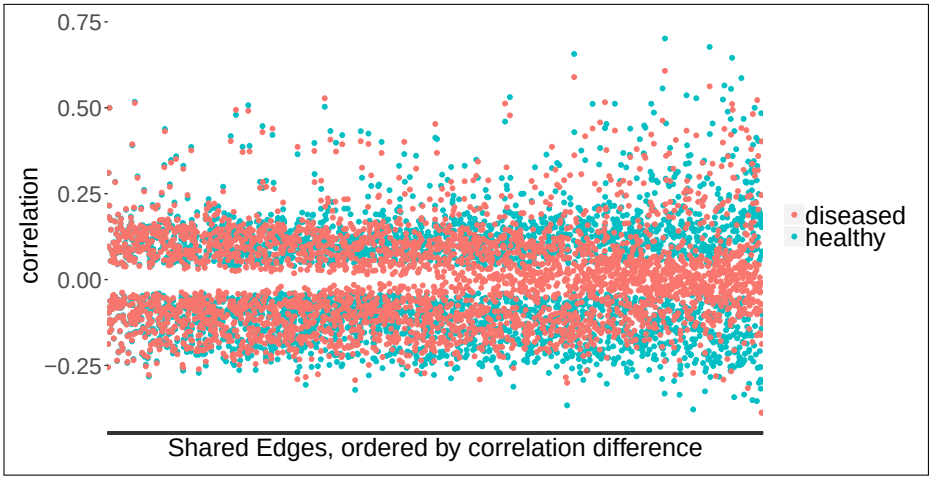
Cohort Data	Unique Edges	Overlapping Edges: Remaining Filter 1	Overlapping Edges: Added from Filter 2	Total Edges: Filter 1 & 2
Healthy	4753	250	34	306
Diseased	4753	162	107	313

Table 5.2: Table including the total overlapping edges before filtering, total overlapping edges in each respective network after the first filter, the overlapping edges that were included from the second filter step, and the total edges resulting from both filtering steps. The last column contains the number of edges identified from the filtering method.

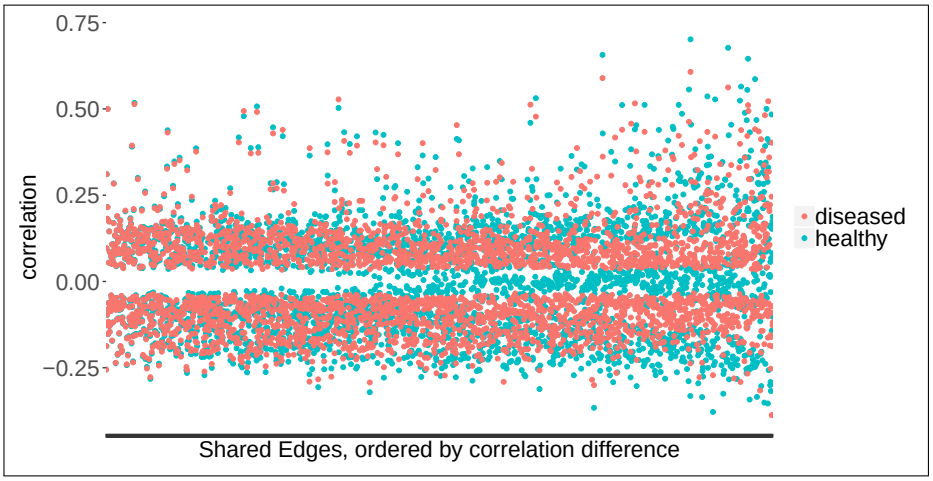
connections between the filtered healthy network and unfiltered diseased network and sorted them by correlation difference as demonstrated in Figure 5.4a. We performed the same method for the filtered diseased network and unfiltered healthy network in Figure 5.4b. In the figures, connections that are the most similar are on the left and those that have the largest difference are on the right. It is interesting to note that as we traverse right toward larger correlation differences the unfiltered network's correlations tend to converge towards 0. Unfortunately we do not compare these parts of the networks if the unfiltered values do not meet the extended filter criteria. Additionally, we do not compare these components because of our use of the NESH score. Recall that it is a value contingent on the overlapping sub-graphs of the two networks. Both Figure 5.4a and 5.4b demonstrate that filtering should be employed to significantly reduce the large amount of noisy data visible in the figures.

When comparing the two sub-figures in Figure 5.4 we saw a significant number of correlations close to 0 in both filtered and unfiltered networks. Following our results in the previous section, we removed a majority of the correlation values that lead to the fully connected matrices despite being close to 0. It appeared that selecting a correlation value threshold of ± 0.25 allowed us to ignore a majority of the low-value correlations.

After experimenting with various extended threshold values, we extended the threshold by a value of 0.1 by considering an extended threshold value of $|c| \geq 0.15$. The implemented version of the normal cutoff value and the extended cutoff for the filtered healthy network is visible in Figure 5.5a. Once again the shared edges are sorted by the correlation difference between the healthy and diseased networks.



(a) Shared edges with the healthy network as the filtered network, and diseased as unfiltered.



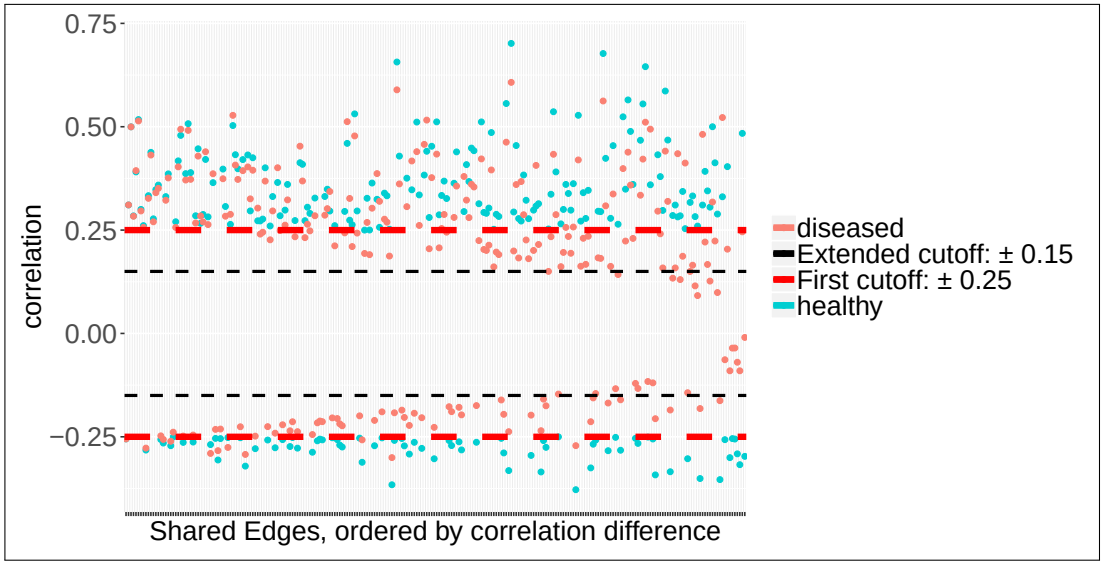
(b) Shared edges with the diseased network as the filtered network, and healthy as unfiltered.

Figure 5.4: All shared connections between the healthy and diseased networks. In each case the filtered network had all correlation values equal to 0 removed, and all edges with p-values greater than 0.05 were removed as well.

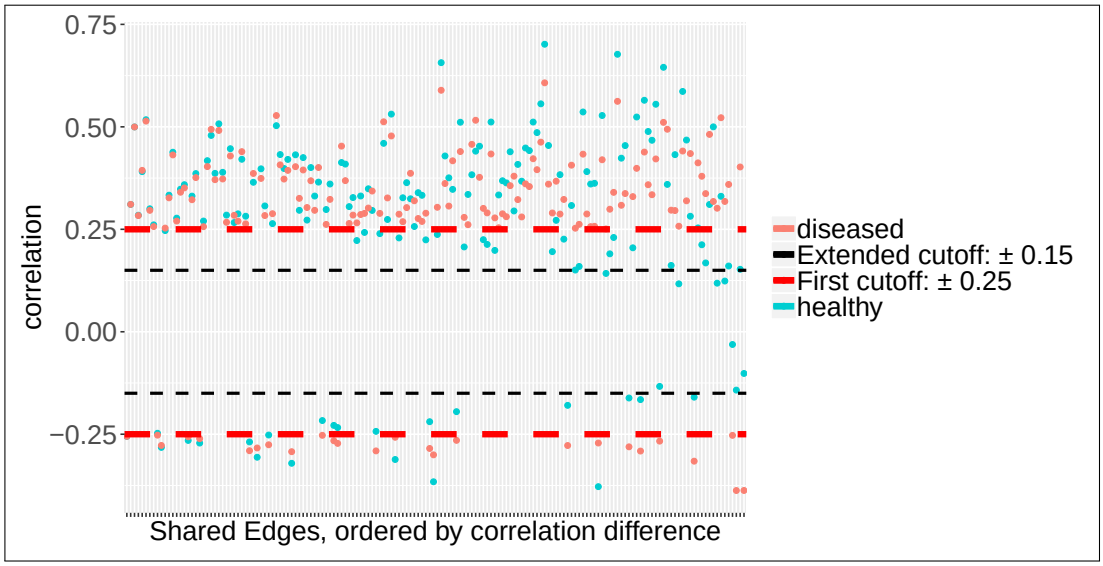
From this plot a majority of the shared edges for the unfiltered diseased network meet the extended cutoff threshold. We use these additional edges in our following analysis for the identification of and comparison of the sub-graphs.

Table 5.2 lists the number of edges associated with this phase of the filtering. For all of the overlapping edges there were 4753 unique edges. Of these there were 250 and 162 remaining after the first filtering step in the healthy and diseased networks respectively. Then after applying the extended correlation threshold, we identified an additional 34 and 107 healthy and diseased respective edges that just missed the initial threshold. These were included in the final networks which ultimately contained a total of 306 and 313 edges respectively.

Using Cytoscape, we generated visual representations of the networks visible in



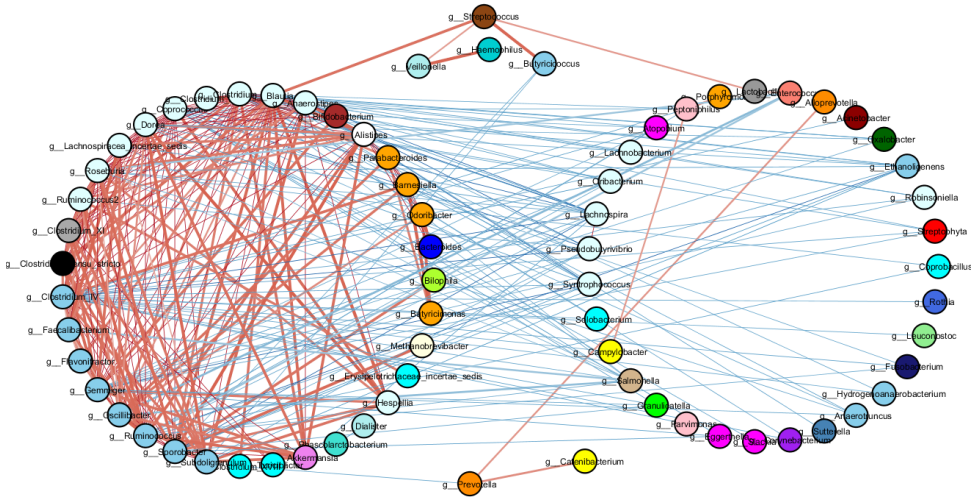
(a) Correlation comparisons for the filtered healthy network and the unfiltered diseased network.



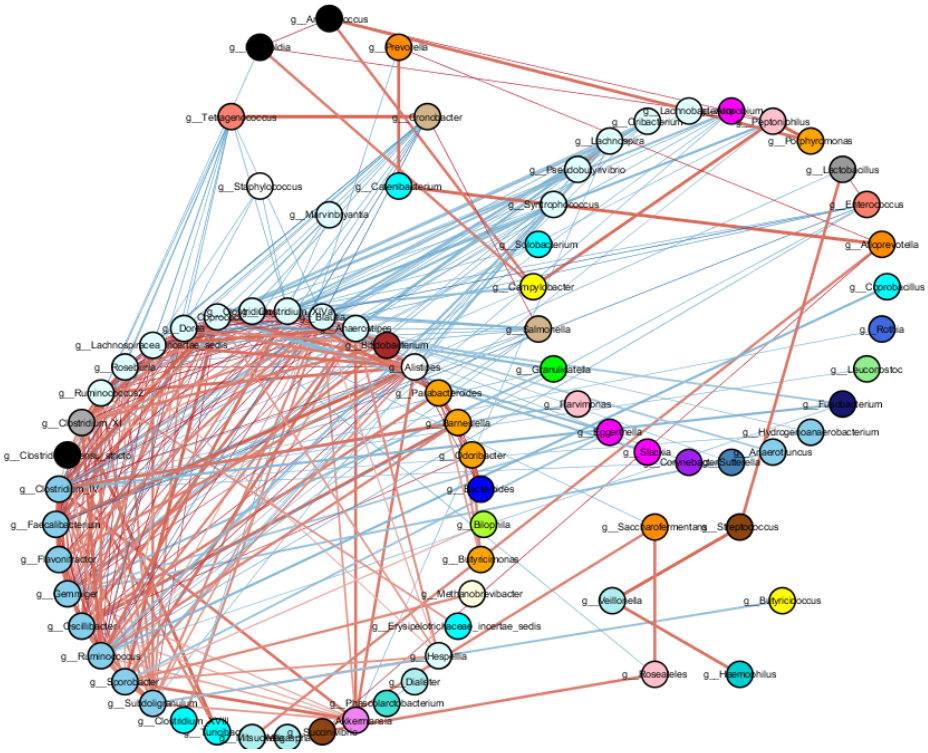
(b) Correlation comparisons for the filtered diseased network and the unfiltered healthy network.

Figure 5.5: Correlation comparisons for the filtered and unfiltered networks. Included in this figure are the correlation cutoff value of 0.25 and the extended correlation threshold of 0.15 which is attributed to an extended correlation value of 0.1.

Figures 5.6a and 5.6b. We colored negative correlations blue, positive correlations red, and scaled the thickness of the connections by the edge weights. We include the figures here so that the reader can visualize the resulting networks. The two networks are similar, but do have different structures related to the exclusive edges present.



(a) Network visualization of the healthy correlation network.



(b) Network visualization of the diseased correlation network.

Figure 5.6: Network visualization of the filtered correlation networks. Red indicates positive correlations, blue indicates negative, and the size of the edges is scaled to the correlation value. Taxa are colored by family, and the grouping of the communities is determined by a hierarchical clustering method.

5.3 Network Statistics

The resulting networks generated from the first and second filtering stages resulted in sparsely connected networks with strong correlations. As was discussed previously,

Network	$\overline{k_i}$	D	l_G	$\overline{C}$	$\overline{g(\mathcal{V})}$	$\overline{C_{nc}}$
Healthy Network (HN)	8.50	0.12	2.44	0.49	53.90	5.43
Diseased Network (DN)	8.03	0.10	2.77	0.49	68.22	4.96
Sub-graph HN	8.48	0.13	2.46	0.52	50.58	5.51
Sub-graph DN	8.03	0.12	2.60	0.51	53.01	5.10

Table 5.3: Table containing the general network statistics for the filtered healthy and diseased networks (HN and DN respectively) as well as their sub-graphs. The statistics are average degree $\overline{k_i}$, graph density D, average path length l_G, average clustering coefficient $\overline{C}$, average betweenness centrality $\overline{g(\mathcal{V})}$, and average coreness centrality $\overline{C_{nc}}$.

the actual network comparisons need to be performed on the overlapping sub-graphs of the two networks. This section presents network statistics of the networks before and after finding the overlapping sub-graphs.

The filtered healthy network had a total of 306 edges which included 72 unique genera. The remaining genera in this network had at least one edge connecting them to another genus. This was a significant reduction in the number of nodes and connections compared to the original FastSpar correlation matrix. Originally there were 107 genera that were fully connected with a total number of 5671 unique edges (as presented in Table 5.1). There was a similar reduction in the genera and correlation values with the diseased network. The filtered disease network had a total of 313 edges including 78 unique genera. Once again there was a significant decrease in the number of edges and genera compared to the original network which contained 6105 edges and 111 unique genera.

The intersection of the two networks revealed that there were 67 genera which were present in both networks. This was roughly a quarter of the original genera as there were in the original raw data sets. The healthy sub-graph contained 284 unique edges and the diseased sub-graph contained 269. Of these, the networks shared 261 edges, with there being 23 unique edges in the healthy sub-graph and 8 unique edges in the diseased sub-graph. Despite the small difference in the unique edges, the general network statistics differ.

Table 5.3 contains some of the general network statistics as described in Section 3.2. Density values for the sub-graphs appear to be what we would expect from the definition of density. The healthy network does have more edges between the genera, and thus the density is larger due to more of the possible edges existing. The biological interpretation of this would be that the higher the density, the more cross-talk there is between microbes in the network. The healthy network also has a smaller average path length which suggests that the additional edges are connected in key parts of the network that reduce multiple paths. So the smaller the value, the more compact the microbial network is. The average clustering coefficient of the two networks is similar, with the healthy network having a value of 0.52 and the diseased network with a value of 0.51. We would expect that the higher the clustering coefficient, the more independently associated microbes will be present in

the network.

The table also contains average values for the degree, betweenness centrality and the coreness centrality. These metrics are important for understanding node level changes in the networks [Kuntal et al., 2018]. While we show the average values of the metrics in Table 5.3, we have included the distribution plots of the values for all of the nodes in the respective networks in the appendix in Figure A.2. It is evident from Figures A.2a, and A.2b that many genera have low degrees and betweenness centrality, but there are several genera with higher values. We would expect higher degrees to be associated with higher importance due to the increased number of direct connections between members of the community. Additionally, higher betweenness centrality highlights the importance as a preferred member of the community. The coreness centrality described in Figure A.2c reveals that there are several genera with a low coreness centrality, and many with higher coreness centrality scores. This indicates that these individuals might have better colony forming capabilities due to the large sub-graphs that can be made with them. We consider these individuals as core areas of the network that are in a core hub community [Kuntal et al., 2018].

After comparing the networks, there are some differences in the taxa associated with the highest degrees, betweenness centrality, and coreness centrality.

We present all of these metrics in the appendix as Table A.1. For the degrees, note that the overlapping networks contain the same top five nodes with respect to degree, even though they have slightly different degree values. This emphasizes the re-wiring that is taking place in the network. If we compare the full networks to the overlapping networks, we find that the healthy network contains four of the same genera, with *g_Gemmiger* and *g_Ruminococcus* substituting for each other. In the diseased network we see that the genera are all conserved, with a reduction in the degrees.

For betweenness centrality three out of the five genera are in the top five between the overlapping networks. *g_Oscillibacter* and *g_Sporobacter* are replaced by *g_Faecalibacterium* and *g_Coprococcus*. Again, this reveals the change in network wiring. Four out of five genera are conserved from the full healthy network to the overlapping healthy network, and the genera are conserved in the diseased networks. Coreness centrality ranks are the same in the overlapping networks, with the values all being 12 and 11 for the healthy and diseased networks respectively. The healthy network has conserved the genera between the full network and the overlapping network, and the diseased network conserves four out of the five genera.

5.4 Network Structure

For the final results we used Kuntal et al.'s web application to visualize community structure and identify the driving nodes in the networks. First we present the visualizations of the re-wiring of the networks with the shuffle plot that is used by Kuntal et al.. The aim of a shuffle plot is to visualize the re-wirings and the

subsequent re-grouping of the nodes in the networks. The NetShift method uses a hierarchical clustering algorithm to identify communities within each network. It then looks across the two networks to determine the re-wiring based upon the nodes and the communities they were clustered into in each respective network. Figure

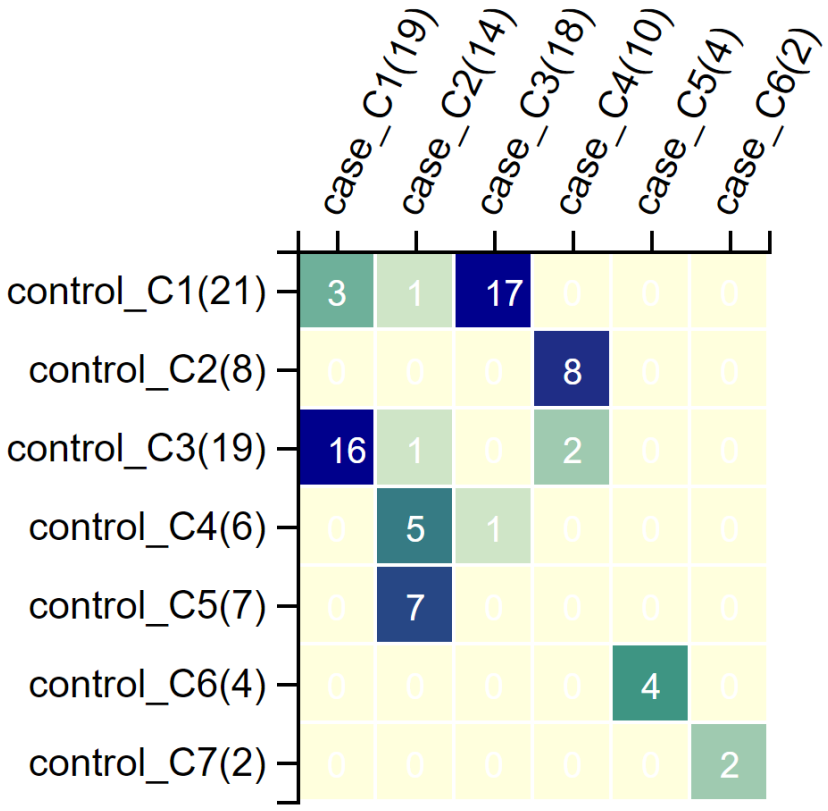


Figure 5.7: NetShift shuffle plot indicating community-level structure changes.

5.7 depicts the community-level structure changes between the healthy (control) and diseased (case) networks. The plot is interpreted by observing the label for either the control or case network, and the given number in parentheses. The given number defines the amount of genera present in the respective community in the respective network. So, for example, of the 21 genera in *case_C2(14)*, 1 is located in *control_C1(21)*, 1 is located in *control_C3(19)*, 5 are located in *control_C4(6)* and 7 are in *control_C5(7)*. Among these genera there is considerable network re-wiring. On the other hand, consider *case_C5(4)*. All of the genera associated with it in the control network stay together in the same community (*control_C6(4)*) in the diseased network. Therefore, the more variance across a row or column, the more different the community structure is.

We have included the network shuffle plot visualization generated from Kuntal et al.'s tool. Considering the shuffle plot in Figure 5.7, the network diagram in Figure 5.8 is used to visualize the healthy community placements of the genera associated with the diseased community *case_C2(14)*. Similarly, Figure 5.9 presents a visualization of the diseased community *case_C5(4)*. In the diagrams we can follow the lines to see how a taxon may shift communities between the healthy and

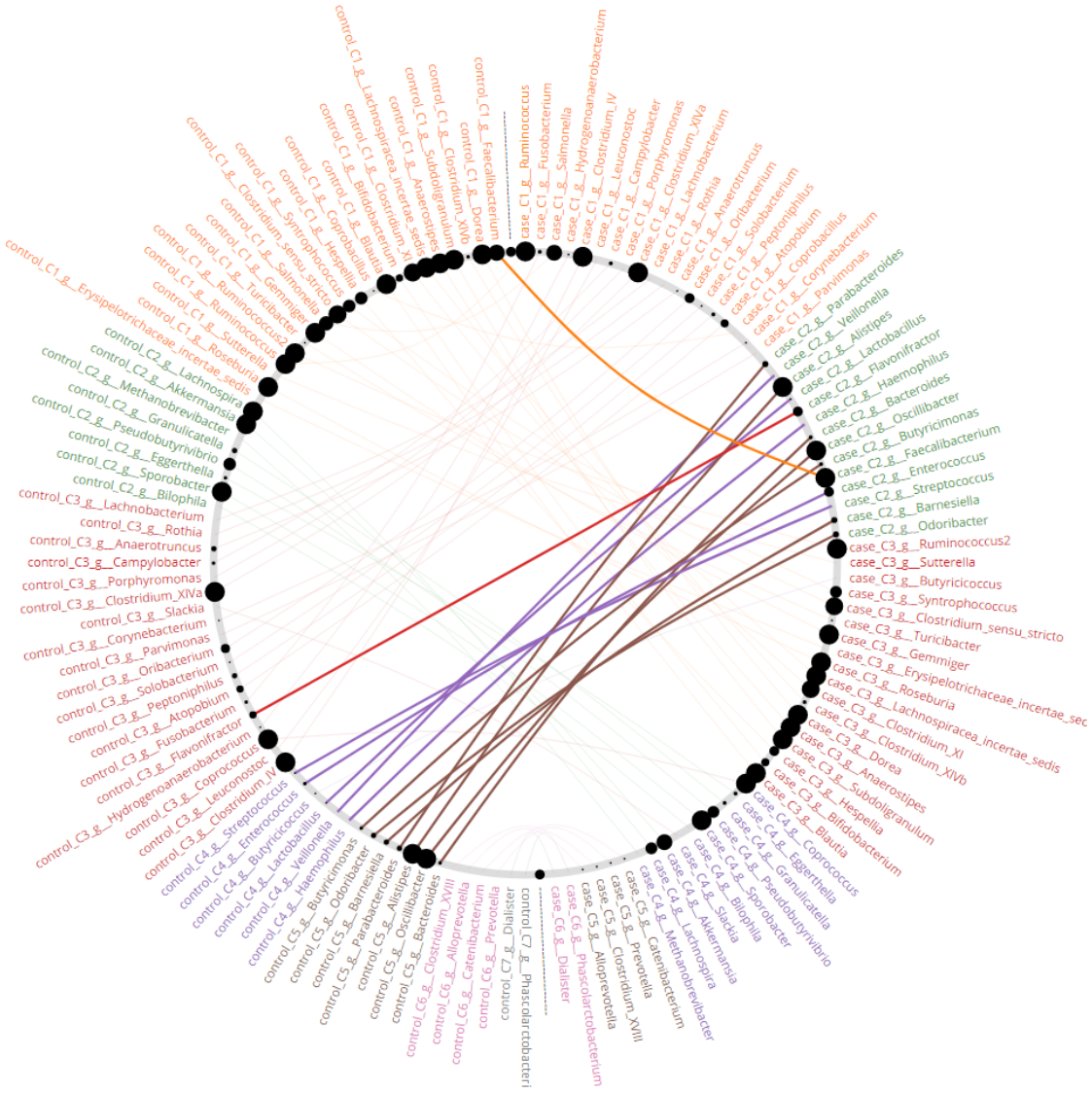


Figure 5.8: The NetShift shuffle diagram highlighting high re-wiring between communities identified by hierarchical clustering. Here we highlight the case community *case_C2(14)* and the variation in the respective nodes diseased communities. All taxa are represented by the control taxa on the left of the diagram, and case taxa on the right. Hierarchical cluster communities are identified by the color of the taxa.

diseased state. The full diagram is included in Figure A.3.

5.5 Network Shift

The Jaccard Edge Index (JEI) of the overlapping sub-graphs was 0.894, and this score avoids the bias that would have been present had we compared the full networks. To take into account the directionality of re-wiring we utilized the NetShift method and identified several “driving” taxa in the networks. These taxa were associated with re-wiring in the network due to their large NESH and betweenness

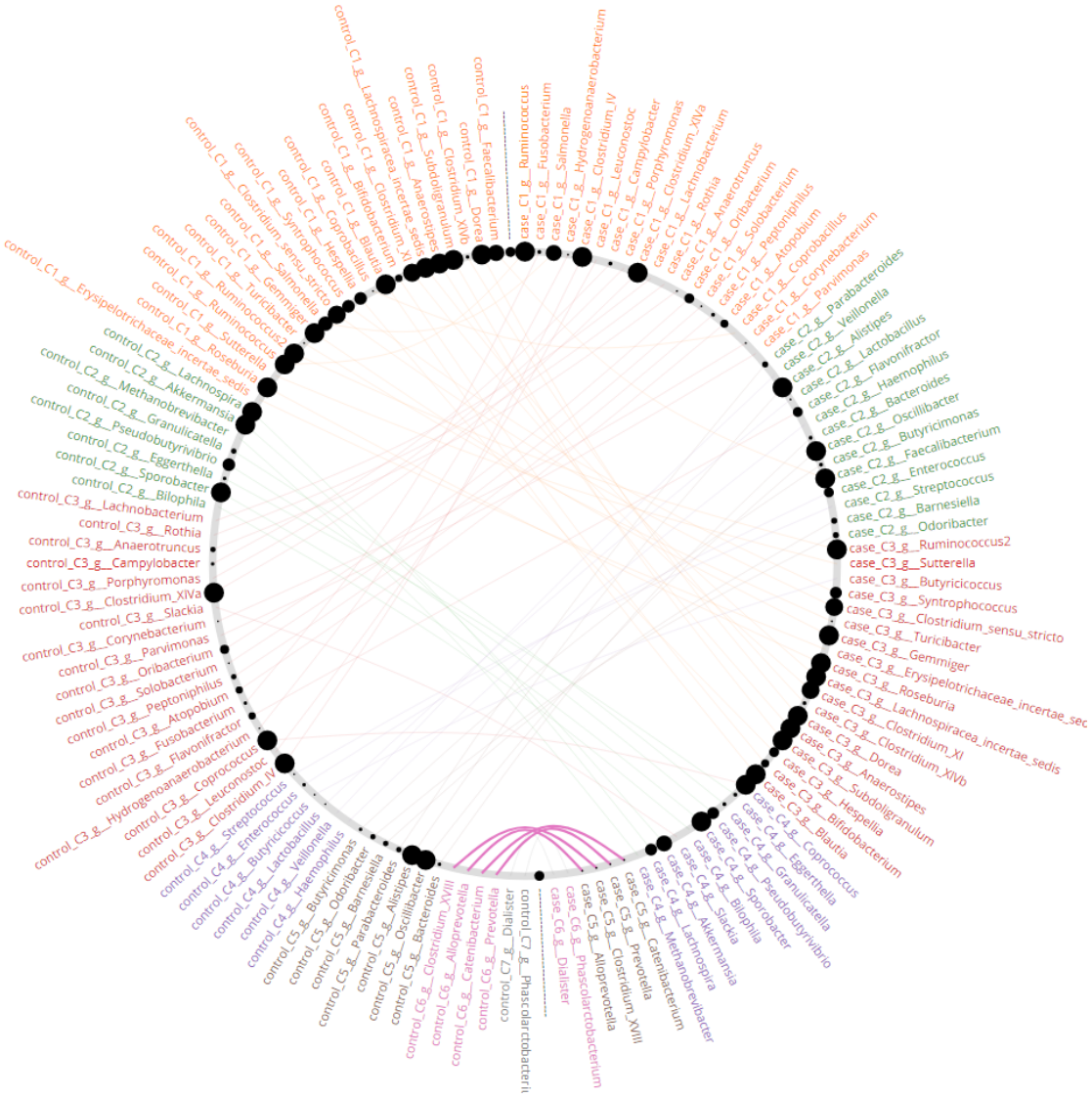


Figure 5.9: The NetShift shuffle diagram highlighting low re-wiring between communities identified by hierarchical clustering. Here we highlight the control community *case_C5(4)* and the conservation of nodes between healthy and diseased. All taxa are represented by the control taxa on the left of the diagram, and case taxa on the right. Hierarchical cluster communities are identified by the color of the taxa.

values. Table 5.4 lists the top 20 genera sorted by their NESH scores. The especially important drivers are identified by higher NESH and betweenness scores from the healthy to the diseased network. Note that it is not required for driving genera to have exclusive edges (as evident for *g_Peptoniphilus*) because of NESH's dependence upon the neighbors of a node. It is also apparent that high NESH scores are not dependent upon high exclusivity or degrees of a node.

We visualized the drivers and edges in the sub-graph networks in Figures 5.10 and 5.11 with Kuntal et al.'s web application. These figures represent diseased exclusive edges highlighted in red and healthy exclusive edges in green. Both plots use the case network's hierarchical clustering for the ordering and coloring of the names of

Genus	Degree (H)	Degree (D)	Exclusive (D)	NESH
g_Alloprevotella	1	2	1	1.029
g_Catenibacterium	1	2	1	1.029
g_Enterococcus*	4	7	3	0.943
g_Lactobacillus*	2	3	1	0.695
g_Butyricicoccus	3	1	0	0.667
g_Anaerotruncus	4	2	0	0.5
g_Coprobacillus	2	1	0	0.5
g_Lachnospira	14	7	0	0.5
g_Streptococcus	4	2	0	0.5
g_Faecalibacterium*	12	15	3	0.486
g_Flavonifractor*	5	6	1	0.362
g_Barnesiella*	6	7	1	0.314
g_Granulicatella	4	3	0	0.25
g_Hespellia	8	6	0	0.25
g_Alistipes*	19	21	2	0.248
g_Dorea*	23	21	1	0.237
g_Sporobacter	23	18	0	0.217
g_Lachnospiracea_inc.*	16	16	1	0.205
g_Akkermansia	12	10	0	0.167
g_Peptoniphilus*	6	5	0	0.167

Table 5.4: Table containing the top 20 genera with the highest NESH scores and their respective degree and edge information. Taxa identified as drivers from the betweenness score and high NESH score are denoted by a * next to their name. Included in the table are the degrees of the respective genus in the healthy (H) and diseased (D) network, the intersect (shared edges), the exclusive edges in the case network, and the NESH score. *g_Lachnospiracea_inc.* was abbreviated from: *g_Lachnospiracea_incertae_sedis*.

the taxa. These plots give further evidence of different structure in the healthy and diseased networks due to the exclusive edges and their influence on the hierarchical clustering and re-wiring of network connections discussed in Section 5.4.

With the nodes being scaled according to a genus's NESH score, and drivers (nodes) colored red, we see the diseased-only edges are primarily connected to driving nodes in Figure 5.10. Additionally, in Figure 5.11, the exclusive healthy edges are connected to many genera that are not defined as drivers. We would expect this result since we are trying to identify taxa that are drivers in the diseased network.

In order for a node to be designated a driver it had to have a high NESH score and a positive delta betweenness. Table 5.5 contains the top 20 NESH score genera and their delta betweenness between the healthy network and the diseased networks. The identified drivers in Table 5.5 all have positive betweenness whereas the genera that were not identified had negative delta betweenness values.

Considering these top 20 NESH scores and delta betweenness values, the Net-

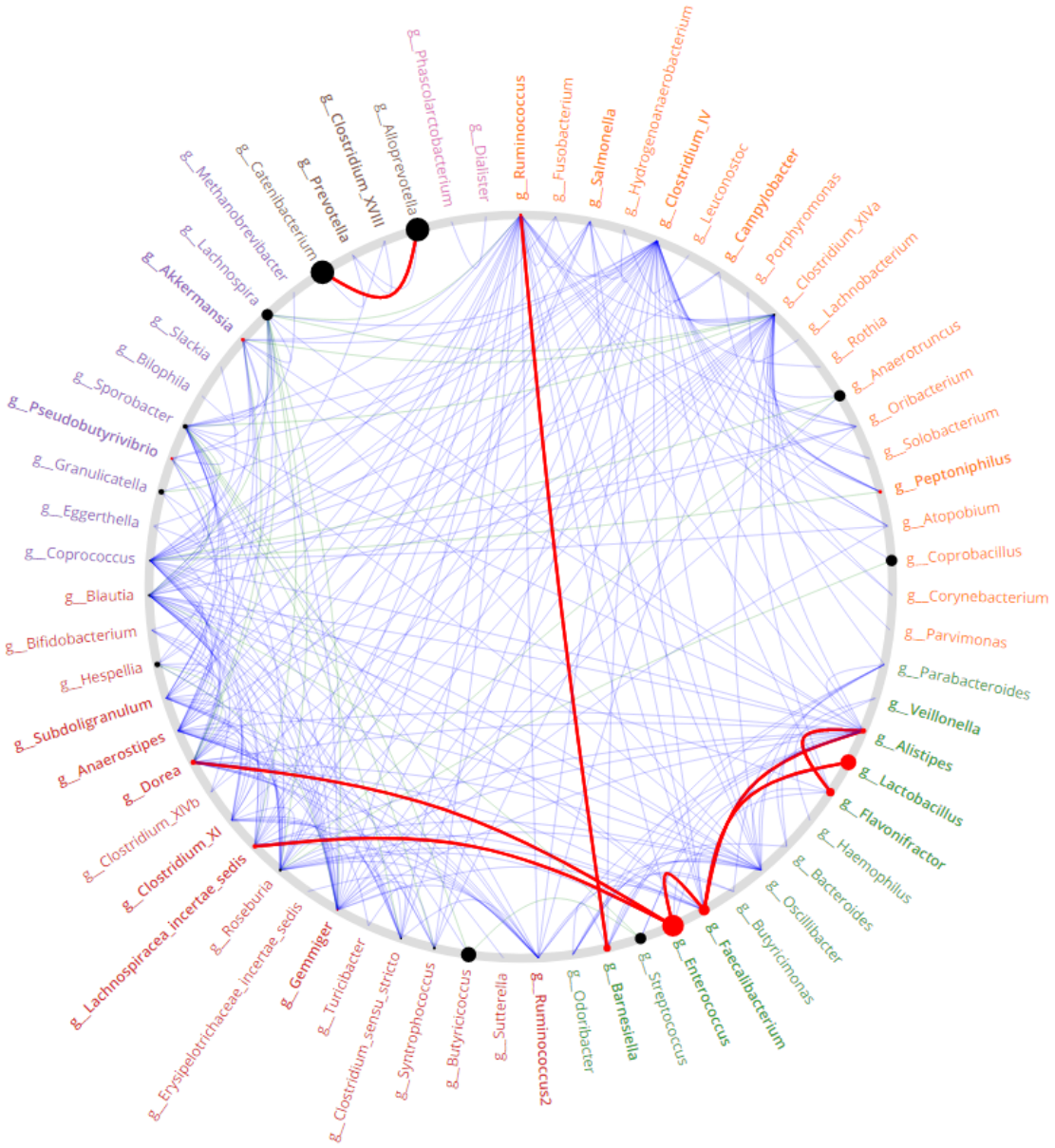


Figure 5.10: NetShift diagram depicting all edges in the sub-graphs with the diseased-only edges highlighted in red. Genera are colored by their assigned sub-communities. Nodes are scaled by their NESH scores, and nodes are colored red if they have been identified as important drivers. Control-only edges are in green, case-only edges are in red, and shared edges are in blue.

Shift methodology identified *g__Enterococcus*, *g__Lactobacillus*, *g__Faecalibacterium*, *g__Flavonifractor*, *g__Barnesiella*, *g__Alistipes*, *g__Dorea*, *g__Peptoniphilus*, and *g__Lachnospiracea__incertae__sedis* as drivers in the diseased state. These metrics suggest that the identified genera play some role in the re-wiring of the network from a healthy to diseased state.

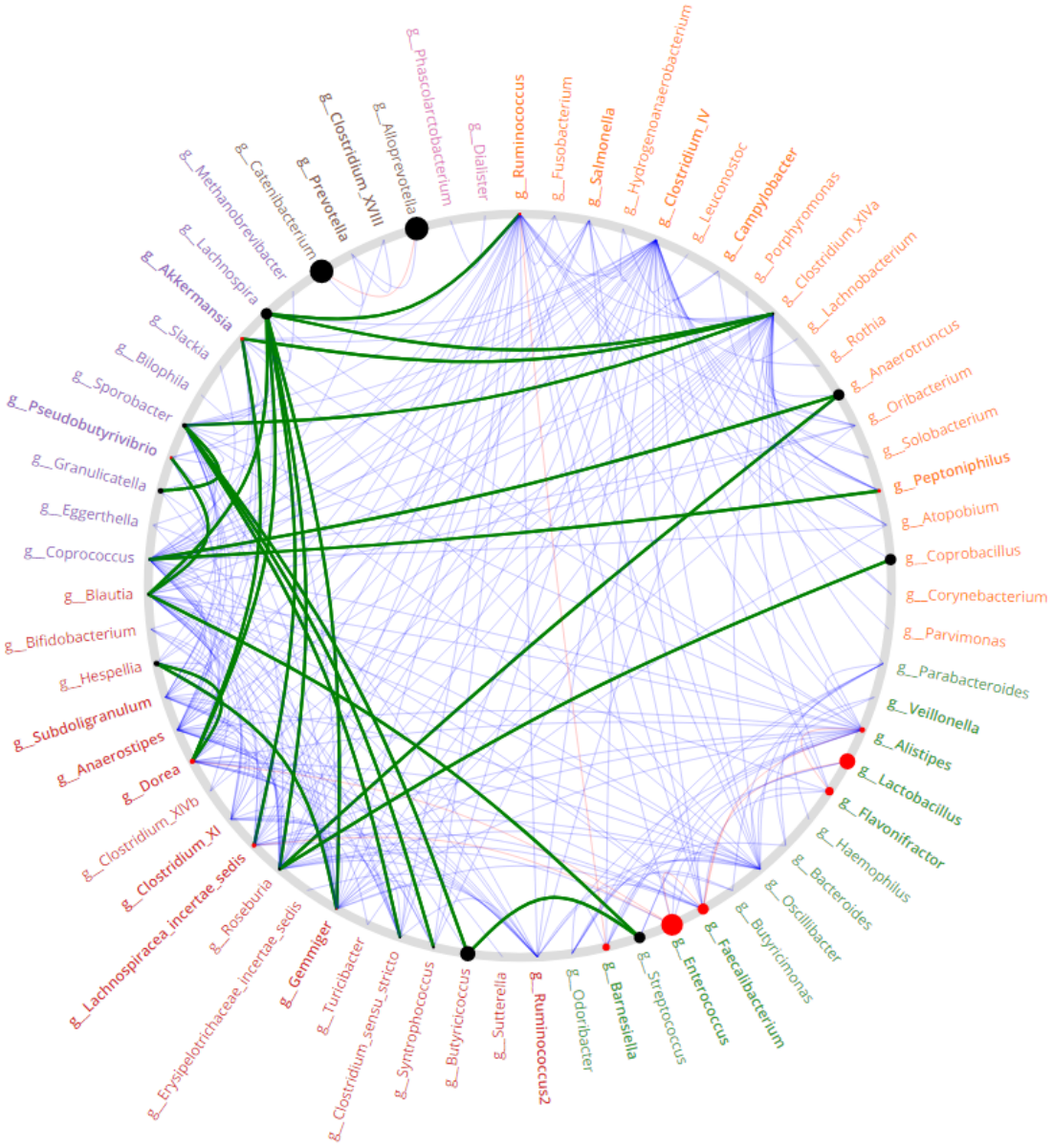


Figure 5.11: NetShift diagram depicting all edges in the sub-graphs with the control-only edges highlighted in green. Genera are colored by their assigned sub-communities. Nodes are scaled by their NESH scores, and nodes are colored red if they have been identified as important drivers. Control-only edges are in green, case-only edges are in red, and shared edges are in blue.

Genus	NESH	Δ Betweenness
g__Alloprevotella	1.029	0
g__Catenibacterium	1.029	0
g__Enterococcus*	0.943	0.005
g__Lactobacillus*	0.695	0.385
g__Butyricicoccus	0.667	-0.018
g__Anaerotruncus	0.5	0
g__Coprobacillus	0.5	0
g__Lachnospira	0.5	-0.003
g__Streptococcus	0.5	-0.046
g__Faecalibacterium*	0.486	0.427
g__Flavonifractor*	0.362	0.001
g__Barnesiella*	0.314	0.018
g__Granulicatella	0.25	0
g__Hespellia	0.25	0
g__Alistipes*	0.248	0.027
g__Dorea*	0.237	0.033
g__Sporobacter	0.217	-0.087
g__Lachnospiracea_incertae_sedis*	0.205	0.019
g__Akkermansia	0.167	0.004
g__Peptoniphilus*	0.167	0.002

Table 5.5: Table containing the identified drivers from the top 20 NESH scores and their related delta betweenness score from control to case. Taxa identified as drivers from the betweenness score and high NESH score are denoted by a * next to their name.

6

Discussion

6.1 Future Work

Our methods followed common practices in the literature, but there still remain many opportunities for future work and exploration.

The careful reader may have noticed that after acquiring and visualizing the filtered correlation networks in Section 5.2, we did not utilize the correlation values further. We used the correlation filtering methods to reduce the connectivity in our original correlation matrices, similarly to how the literature tends to employ arbitrary cutoff values to visualize these networks. The resulting networks should contain the structural information that we were curious about. Our time was limited to investigate comparative techniques for similar networks based upon correlations, but we expect that there are relevant methods in the literature.

We decided to employ the NetShift method to identify drivers in the network, but the Kuntal et al. [2018] implementation utilizes undirected and unweighted networks. Further work should focus on modifying the NetShift method or NESH metric to include correlation value weights. The additional information here could further quantify changes in correlation values and connections between neighbors. This type of comparison would need to consider potential bias arising from the different networks' correlation values potentially not being comparable.

This study made use of publicly available meta-analysis data, and we split the data into respective healthy and diseased networks. While a general investigation between healthy and diseased gut microbiomes is useful for identifying genera that are drivers in dysbiosis, it would be wise to perform comparisons between healthy guts and individual diseases. We did not perform this type of analysis to identify driving microbes for specific diseases because of the small number of data points for some diseases. We recommend that this comparison be performed for a diseases where there is a sufficiently large sample size. In this case, the methods we followed should be used to build correlation networks for all of the specific diseases. Then the NetShift method could be applied between the healthy correlation network and the respective diseased networks. Results here could shed light on whether disease type (metabolic, cardiovascular, etc.) is associated with certain driving taxa, or whether all diseased guts are similar.

The development of more robust HT technologies will be of use for future imple-

mentations of these methods. For example, publicly available shallow shotgun sequencing will be easier to combine for meta-analysis work and could allow for this analysis to be performed at the species level. Species level resolution will be vital to researchers aiming to understand the gut microbiome. Additionally, after investigating the Duvallet et al. [2017] data we were unable to build a multiclass classification model to identify specific disease states in the data. Future technologies should be able to give better resolution of microbiome structure, and we expect that with enough samples, the multiclass problem could be solved.

All of the data in this study comes from a single stool sample from each patient. While we have been able to discern structure in healthy and diseased microbiome networks from this single snapshot, the data does not necessarily represent all of the possible states that these networks could reside in. We know that the gut microbiome has a dynamic community, so temporal data from longitudinal studies would be very useful in understanding both the variability in health states as well as the variability in a given person's microbiome. Such data will contribute to a better understanding of the dynamics in the gut and specifically allow for the identification of driving taxa or key community members.

Identifying structure and key drivers in these microbial networks is just a starting point. Once found it will be useful to begin modeling actual microbe-microbe metabolic interactions. Magnúsdóttir et al. [2016] are currently reconstructing the metabolic pathways for microbes in the gut. With their reconstructions it is possible to model the dynamics of these communities through flux balance and co-growth simulations to predict how communities change over time. It is also useful for identifying microbe-microbe relationships, and the Magnúsdóttir et al. work is already being used for such simulations[1]. An extension of this will be to investigate how metabolites produced in the gut impact the human host, and how metabolites from the human host will impact the gut. This type of research is going to be useful for the development of gut-based microbiome therapies and will expedite the research and development process in the wet-lab.

6.2 Societal and Ethical Aspects

Human microbiome research is very reliant on real patient data. Because of this reliance there are many regulations in place to ensure the protection of the identity of patients and their medical history. Such studies require the patient to indicate the scope of their consent for data usage. Some patients may opt for their data to only be used in a specific study, while others are open to their data being used across multiple types of studies. It is up to the life sciences, healthcare, and academic industries to follow these regulations and patient consent to avoid misuse of data.

If data is not anonymized, patients could unwillingly have their personal information and medical history leaked to the public. Such a blatant violation of individual's

[1] `https://opencobra.github.io/cobratoolbox/stable/tutorials/tutorialMicrobeMicrobeInteractions.html`

privacy is often times against the law, and could negatively impact a patient's future. Research organizations go through many steps to ensure that data is only accessed within the scope of consent. Often times private patient data is stored in databases that log all accessions from researchers in the organization. This is done to track the usage of data and to identify individuals who may use the data maliciously.

The data in this thesis comes from an aggregation of 28 studies which were all followed standard medical regulations of clinical data. While there was associated metadata available for each gut microbiome sample, the different labs anonymized their patient information and followed the scope of their patients' consent. Further research should follow the regulations put in place by governing health bodies, and researcher should be sure to be aware of the societal and personal impacts of misuse of human gut microbiome data.

6.3 Conclusion

In this thesis we used different statistical and graph theoretic methods to identify nodes that are closely associated with overall structure change between healthy and diseased networks. After filtering our raw OTU data so that taxa occured in at least 5% of our samples, we estimated fractional abundances for the taxa in each sample by drawing from a Dirichilet distribution. The resulting correlation networks generated from the fractional abundances were then filtered under the assumption that the network in the gut is sparse. From here we implemented an extended filtering threshold in order to reduce artifacts that could be introduced by a basic cutoff value. Then the resulting networks were compared using the NetShift methodology and we identified 9 "driving" taxa that contributed to the re-wiring of a healthy to diseased gut microbiome.

These findings highlight the structural diversity of real-world networks and the need for new theoretical explanations of these non-scale-free patterns [Broido and Clauset, 2019]. In this thesis, systems and computational biology play a role in understanding the behavior of the gut microbiota. However, current knowledge of the human gut microbiome is still in its infancy. There still exists a large need for further research into microbial interactions and behaviors in the gut. Development of HT technologies will aid many aspects of microbiome research. Particularly relevant influences will be on metagenomic resolution as well as further research into metabolic pathways, microbe-microbe interactions, and microbe-host interactions. Together this research should uncover mechanisms of action relating to community stability in the gut microbiome and will potentially allow for the development of treatments for a multitude of diseases.

Bibliography

[1] Alan Agresti and David B. Hitchcock. Bayesian Inference for Categorical Data Analysis. *Statistical Methods and Applications*, 14:297–330, 12 2005. doi: 10.1007/s10260-005-0121-y.

[2] John Aitchison. A Concise Guide to Compositional Data Analysis. In *2nd Compositional Data Analysis Workshop*. 2nd Compositional Data Analysis Workshop; Girona, Italy, 2003. URL `http://tiny.cc/s06f5y`.

[3] John Aitchison. *The Statistical Analysis of Compositional Data*. The Blackburn Press, 2003. ISBN 1930665784.

[4] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Internet: Diameter of the world-wide web. *Nature*, 401:130–131, 1999.

[5] Aimon K. Alkanani, Naoko Hara, Peter A. Gottlieb, Diana Ir, Charles E. Robertson, Brandie D. Wagner, Daniel N. Frank, and Danny Zipris. Alterations in Intestinal Microbiota Correlate With Susceptibility to Type 1 Diabetes. *Diabetes*, 64(10):3510–3520, June 2015. doi: 10.2337/db14-1847. URL `https://doi.org/10.2337/db14-1847`.

[6] Iain Anderson, Markus Göker, Matt Nolan, Susan Lucas, Nancy Hammon, Shweta Deshpande, Jan-Fang Cheng, Roxanne Tapia, Cliff Han, Lynne Goodwin, Sam Pitluck, Marcel Huntemann, Konstantinos Liolios, Natalia Ivanova, Ioanna Pagani, Konstantinos Mavromatis, Galina Ovchinikova, Amrita Pati, Amy Chen, Krishna Palaniappan, Miriam Land, Loren Hauser, Evelyne-Marie Brambilla, Harald Huber, Montri Yasawong, Manfred Rohde, Stefan Spring, Birte Abt, Johannes Sikorski, Reinhard Wirth, John C. Detter, Tanja Woyke, James Bristow, Jonathan A. Eisen, Victor Markowitz, Philip Hugenholtz, Nikos C. Kyrpides, Hans-Peter Klenk, and Alla Lapidus. Complete genome sequence of the hyperthermophilic chemolithoautotroph Pyrolobus fumarii type strain (1AT). *Standards in Genomic Sciences*, 4(3):381–392, July 2011. doi: 10.4056/sigs.2014648. URL `https://doi.org/10.4056/sigs.2014648`.

[7] Joonhyun Bae and Sangwook Kim. Identifying and ranking influential spreaders in complex networks by neighborhood coreness. *Physica A: Statistical Mechanics and its Applications*, 395:549–559, February 2014. doi: 10.1016/j.physa.2013.10.047. URL `https://doi.org/10.1016/j.physa.2013.10.047`.

[8] Albert-László Barabási. Love is All You Need: Clauset's fruitless search for scale-free networks. Web, March 2018. URL `https://web.archive.org/web/20190426011850/https://www.barabasilab.com/post/love-is-all-you-need`.

[9] Albert-László Barabási and Reka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999. doi: 10.1126/science.286.5439.509. URL `http://www.sciencemag.org/cgi/content/abstract/286/5439/509`.

[10] Albert-László Barabási, Réka Albert, and Hawoong Jeong. Mean-field theory for scale-free random networks. *Physica A: Statistical Mechanics and its Applications*, 272(1-2):173–187, 1999. ISSN 0378-4371. doi: 10.1016/S0378-4371(99)00291-5.

[11] Albert-László Barabási and Márton Pósfai. *Network Science*. Cambridge University Press, Cambridge, 2016. ISBN 9781107076266 1107076269. URL http://barabasi.com/networksciencebook/.

[12] Albert Batushansky, David Toubiana, and Aaron Fait. Correlation-Based Network Generation, Visualization, and Analysis as a Powerful Tool in Biological Studies: A Case Study in Cancer Cell Metabolism. *BioMed Research International*, 2016:1–9, 2016. doi: 10.1155/2016/8313272. URL https://doi.org/10.1155/2016/8313272.

[13] Nielson T. Baxter, Mack T. Ruffin, Mary A. M. Rogers, and Patrick D. Schloss. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome Medicine*, 8(1), April 2016. doi: 10.1186/s13073-016-0290-3. URL https://doi.org/10.1186/s13073-016-0290-3.

[14] Anna D. Broido and Aaron Clauset. Scale-free networks are rare. *Nature Communications*, 10(1), March 2019. doi: 10.1038/s41467-019-08746-5. URL https://doi.org/10.1038/s41467-019-08746-5.

[15] A. Buccianti, G. Mateu-Figueras, V. Pawlowsky-Glahn, and Editors. *Compositional Data Analysis in the Geosciences: From Theory to Practice - Special Publication no 264 (Geological Society Special Publication)*. Geological Society of London, 2006. ISBN 1862392056.

[16] Weiguang Chen, Fanlong Liu, Zongxin Ling, Xiaojuan Tong, and Charlie Xiang. Human Intestinal Lumen and Mucosa-Associated Microbiota in Patients with Colorectal Cancer. *PLoS ONE*, 7(6):e39743, June 2012. doi: 10.1371/journal.pone.0039743. URL https://doi.org/10.1371/journal.pone.0039743.

[17] E. K. Costello, K. Stagaman, L. Dethlefsen, B. J. M. Bohannan, and D. A. Relman. The Application of Ecological Theory Toward an Understanding of the Human Microbiome. *Science*, 336(6086):1255–1262, June 2012. doi: 10.1126/science.1224203. URL https://doi.org/10.1126/science.1224203.

[18] Duy M. Dinh, Gretchen E. Volpe, Chad Duffalo, Seema Bhalchandra, Albert K. Tai, Anne V. Kane, Christine A. Wanke, and Honorine D. Ward. Intestinal Microbiota, Microbial Translocation, and Systemic Inflammation in Chronic HIV Infection. *Journal of Infectious Diseases*, 211(1):19–27, July 2014. doi: 10.1093/infdis/jiu409. URL https://doi.org/10.1093/infdis/jiu409.

[19] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes. k-Core Organization of Complex Networks. *Phys. Rev. Lett.*, 96:040601, Feb 2006. doi: 10.1103/PhysRevLett.96.040601. URL https://link.aps.org/doi/10.1103/PhysRevLett.96.040601.

[20] Claire Duvallet. microbiomeHD: Cross-disease comparison of case-control gut microbiome studies. GitHub Repository, May 2018. URL https://github.com/cduvallet/microbiomeHD. original-date: 2016-10-20T03:15:56Z.

[21] Claire Duvallet, Sean M. Gibbons, Thomas Gurry, Rafael A. Irizarry, and Eric J. Alm. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nature Communications*, 8(1), December 2017. ISSN 2041-1723. doi: 10.1038/s41467-017-01973-8. URL http://www.nature.com/articles/s41467-017-01973-8.

[22] P. Erdös and A. Rényi. On random graphs i. *Publicationes Mathematicae Debrecen*, 6:290, 1959.

[23] Leonhard Euler. Solutio problematis ad geometriam situs pertinentis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae*, 8:128–140, 1736.

[24] Richard Foote. Mathematics and Complex Systems. *Science*, 318(5849):410–412, October 2007. doi: 10.1126/science.1141754. URL https://doi.org/10.1126/science.1141754.

[25] Linton C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40 (1):35, March 1977. doi: 10.2307/3033543. URL https://doi.org/10.2307/3033543.

[26] Jonathan Friedman and Eric J. Alm. Inferring correlation networks from genomic survey data. *PLoS Computational Biology*, 8(9):e1002687, September 2012. doi: 10.1371/journal.pcbi.1002687. URL https://doi.org/10.1371/journal.pcbi.1002687.

[27] John D. Fulton. Microorganisms of the Upper Atmosphere IV. Microorganisms of a Land Air Mass as It Traverses an Ocean. *Applied Microbiology*, 14:241–244, March 1966.

[28] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis (Chapman & Hall/CRC Texts in Statistical Science)*. Chapman and Hall/CRC, nov 2013. ISBN 1439840954. URL https://www.xarg.org/ref/a/1439840954/.

[29] Dirk Gevers, Subra Kugathasan, Lee A. Denson, Yoshiki Vázquez-Baeza, Will Van Treuren, Boyu Ren, Emma Schwager, Dan Knights, Se Jin Song, Moran Yassour, Xochitl C. Morgan, Aleksandar D. Kostic, Chengwei Luo, Antonio González, Daniel McDonald, Yael Haberman, Thomas Walters, Susan Baker, Joel Rosh, Michael Stephens, Melvin Heyman, James Markowitz, Robert Baldassano, Anne Griffiths, Francisco Sylvester, David Mack, Sandra Kim, Wallace Crandall, Jeffrey Hyams, Curtis Huttenhower, Rob Knight, and Ramnik J. Xavier. The Treatment-Naive Microbiome in New-Onset Crohn's Disease. *Cell Host & Microbe*, 15(3):382–392, March 2014. doi: 10.1016/j.chom.2014.02.005. URL https://doi.org/10.1016/j.chom.2014.02.005.

[30] Bogdan Giuşcă. The problem of the seven bridges of königsberg. CC BY-SA 3.0, 2005. URL https://commons.wikimedia.org/wiki/File:Konigsberg_bridges.png. File: Konigsberg bridges.png.

[31] Joshua E. Goldford, Nanxi Lu, Djordje Bajić, Sylvie Estrela, Mikhail Tikhonov, Alicia Sanchez-Gorostiaga, Daniel Segrè, Pankaj Mehta, and Alvaro Sanchez. Emergent simplicity in microbial community assembly. *Science*, 361(6401):469–474, 2018. ISSN 0036-8075. doi: 10.1126/science.aat1168. URL http://science.sciencemag.org/content/361/6401/469.

[32] Julia K. Goodrich, Jillian L. Waters, Angela C. Poole, Jessica L. Sutter, Omry Koren, Ran Blekhman, Michelle Beaumont, William Van Treuren, Rob Knight, Jordana T. Bell, Timothy D. Spector, Andrew G. Clark, and Ruth E. Ley. Human Genetics Shape the Gut Microbiome. *Cell*, 159(4):789–799, November 2014. doi: 10.1016/j.cell.2014.09.053. URL https://doi.org/10.1016/j.cell.2014.09.053.

[33] B.C. Goodwin. *Temporal Organization in Cells; A Dynamic Theory of Cellular Control Process*. London: Academic Press, 1963.

[34] Stéphane Hacquard, Ruben Garrido-Oter, Antonio González, Stijn Spaepen, Gail Ackermann, Sarah Lebeis, Alice C. McHardy, Jeffrey L. Dangl, Rob Knight, Ruth Ley, and Paul Schulze-Lefert. Microbiota and Host Nutrition across Plant and Animal Kingdoms. *Cell Host & Microbe*, 17(5):603–616, May 2015. doi: 10.1016/j.chom.2015.04.009. URL https://doi.org/10.1016/j.chom.2015.04.009.

[35] Petter Holme. Rare and everywhere: Perspectives on scale-free networks. *Nature Communications*, 10(1), March 2019. doi: 10.1038/s41467-019-09038-8. URL https://doi.org/10.1038/s41467-019-09038-8.

[36] Dae-Wook Kang, Jin Gyoon Park, Zehra Esra Ilhan, Garrick Wallstrom, Joshua LaBaer, James B. Adams, and Rosa Krajmalnik-Brown. Reduced Incidence of Prevotella and Other Fermenters in Intestinal Microflora of Autistic Children. *PLoS ONE*, 8(7):e68322, July 2013. doi: 10.1371/journal.pone.0068322. URL https://doi.org/10.1371/journal.pone.0068322.

[37] Hiroaki Kitano. Systems Biology: A Brief Overview. *Science*, 295(5560):1662–1664, March 2002. doi: 10.1126/science.1069492. URL https://doi.org/10.1126/science.1069492.

[38] Rob Knight, Alison Vrbanac, Bryn C. Taylor, Alexander Aksenov, Chris Callewaert, Justine Debelius, Antonio Gonzalez, Tomasz Kosciolek, Laura-Isobel McCall, Daniel McDonald, Alexey V. Melnik, James T. Morton, Jose Navas, Robert A. Quinn, Jon G. Sanders, Austin D. Swafford, Luke R. Thompson, Anupriya Tripathi, Zhenjiang Z. Xu, Jesse R. Zaneveld, Qiyun Zhu, J. Gregory Caporaso, and Pieter C. Dorrestein. Best practices for analysing microbiomes. *Nature Reviews Microbiology*, 16(7):410–422, July 2018. ISSN 1740-1526, 1740-1534. doi: 10.1038/s41579-018-0029-9. URL http://www.nature.com/articles/s41579-018-0029-9.

[39] Bhusan K. Kuntal, Pranjal Chandrakar, Sudipta Sadhu, and Sharmila S. Mande. 'NetShift': a methodology for understanding 'driver microbes' from healthy and disease microbiome datasets. *The ISME Journal*, 13(2):442–454, October 2018. doi: 10.1038/s41396-018-0291-x. URL https://doi.org/10.1038/s41396-018-0291-x.

[40] Nick Lane. The unseen world: reflections on Leeuwenhoek (1677) 'Concerning little animals'. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1666): 20140344–20140344, March 2015. doi: 10.1098/rstb.2014.0344. URL https://doi.org/10.1098/rstb.2014.0344.

[41] Mehdi Layeghifard, David M. Hwang, and David S. Guttman. Disentangling Interactions in the Microbiome: A Network Perspective. *Trends in Microbiology*, 25(3):217–228, March 2017. doi: 10.1016/j.tim.2016.11.008. URL https://doi.org/10.1016/j.tim.2016.11.008.

[42] Jiayu Lin. On The Dirichlet Distribution. M.Sc. Thesis, Queen's University, Kingston, Ontario, Canada, September 2016. URL https://mast.queensu.ca/~communications/Papers/msc-jiayu-lin.pdf.

[43] Catherine A. Lozupone, Marcella Li, Thomas B. Campbell, Sonia C. Flores, Derek Linderman, Matthew J. Gebert, Rob Knight, Andrew P. Fontenot, and Brent E. Palmer. Alterations in the Gut Microbiota Associated with HIV-1 Infection. *Cell Host & Microbe*, 14 (3):329–339, September 2013. doi: 10.1016/j.chom.2013.08.006. URL https://doi.org/10.1016/j.chom.2013.08.006.

[44] Linyuan Lü, Tao Zhou, Qian-Ming Zhang, and H. Eugene Stanley. The H-index of a network node and its relation to degree and coreness. *Nature Communications*, 7(1), January 2016. doi: 10.1038/ncomms10168. URL https://doi.org/10.1038/ncomms10168.

[45] Stefanía Magnúsdóttir, Almut Heinken, Laura Kutt, Dmitry A Ravcheev, Eugen Bauer, Alberto Noronha, Kacy Greenhalgh, Christian Jäger, Joanna Baginska, Paul Wilmes, Ronan M T Fleming, and Ines Thiele. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nature Biotechnology*, 35(1):81–89, November 2016. ISSN 1087-0156, 1546-1696. doi: 10.1038/nbt.3703. URL http://www.nature.com/doifinder/10.1038/nbt.3703.

[46] Siddhartha Mandal, Will Van Treuren, Richard A. White, Merete Eggesbø, Rob Knight, and Shyamal D. Peddada. Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial Ecology in Health & Disease*, 26(0), May 2015. doi: 10.3402/mehd.v26.27663. URL https://doi.org/10.3402/mehd.v26.27663.

[47] Julian R. Marchesi and Jacques Ravel. The vocabulary of microbiome research: a proposal. *Microbiome*, 3(1), July 2015. doi: 10.1186/s40168-015-0094-5. URL https://doi.org/10.1186/s40168-015-0094-5.

[48] María Esther Mejía-León, Joseph F. Petrosino, Nadim Jose Ajami, María Gloria Domínguez-Bello, and Ana María Calderón de la Barca. Fecal microbiota imbalance in

Mexican children with type 1 diabetes. *Scientific Reports*, 4(1), January 2014. doi: 10.1038/srep03814. URL https://doi.org/10.1038/srep03814.

[49] Ahmed M'hamdi and Mohamed Nemiche. Bottom-Up and Top-Down Approaches to Simulate Complex Social Phenomena. *International Journal of Applied Evolutionary Computation*, 9(2):1–16, April 2018. doi: 10.4018/ijaec.2018040101. URL https://doi.org/10.4018/ijaec.2018040101.

[50] Xochitl C. Morgan, Timothy L. Tickle, Harry Sokol, Dirk Gevers, Kathryn L. Devaney, Doyle V. Ward, Joshua A. Reyes, Samir A. Shah, Neal LeLeiko, Scott B. Snapper, Athos Bousvaros, Joshua Korzenik, Bruce E. Sands, Ramnik J. Xavier, and Curtis Huttenhower. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biology*, 13(9):R79, 2012. doi: 10.1186/gb-2012-13-9-r79. URL https://doi.org/10.1186/gb-2012-13-9-r79.

[51] D. R. Nemergut, S. K. Schmidt, T. Fukami, S. P. O'Neill, T. M. Bilinski, L. F. Stanish, J. E. Knelman, J. L. Darcy, R. C. Lynch, P. Wickey, and S. Ferrenberg. Patterns and Processes of Microbial Community Assembly. *Microbiology and Molecular Biology Reviews*, 77(3):342–356, September 2013. doi: 10.1128/mmbr.00051-12. URL https://doi.org/10.1128/mmbr.00051-12.

[52] Denis Noble. The rise of computational biology. *Nature Reviews Molecular Cell Biology*, 3 (6):459–463, June 2002. doi: 10.1038/nrm810. URL https://doi.org/10.1038/nrm810.

[53] Marc Noguera-Julian, Muntsa Rocafort, Yolanda Guillén, Javier Rivera, Maria Casadellà, Piotr Nowak, Falk Hildebrand, Georg Zeller, Mariona Parera, Rocío Bellido, Cristina Rodríguez, Jorge Carrillo, Beatriz Mothe, Josep Coll, Isabel Bravo, Carla Estany, Cristina Herrero, Jorge Saz, Guillem Sirera, Ariadna Torrela, Jordi Navarro, Manel Crespo, Christian Brander, Eugènia Negredo, Julià Blanco, Francisco Guarner, Maria Luz Calle, Peer Bork, Anders Sönnerborg, Bonaventura Clotet, and Roger Paredes. Gut Microbiota Linked to Sexual Preference and HIV Infection. *EBioMedicine*, 5:135–146, March 2016. doi: 10.1016/j.ebiom.2016.01.032. URL https://doi.org/10.1016/j.ebiom.2016.01.032.

[54] Bernhard Palsson. The challenges of in silico biology. *Nature Biotechnology*, 18(11): 1147–1150, November 2000. doi: 10.1038/81125. URL https://doi.org/10.1038/81125.

[55] Bernhard Ø. Palsson. *Systems Biology: Properties of Reconstructed Networks*. Cambridge University Press, 2006. doi: 10.1017/CBO9780511790515.

[56] Eliseo Papa, Michael Docktor, Christopher Smillie, Sarah Weber, Sarah P. Preheim, Dirk Gevers, Georgia Giannoukos, Dawn Ciulla, Diana Tabbaa, Jay Ingram, David B. Schauer, Doyle V. Ward, Joshua R. Korzenik, Ramnik J. Xavier, Athos Bousvaros, and Eric J. Alm. Non-Invasive Mapping of the Gastrointestinal Microbiota Identifies Children with Inflammatory Bowel Disease. *PLoS ONE*, 7(6):e39242, June 2012. doi: 10.1371/journal.pone.0039242. URL https://doi.org/10.1371/journal.pone.0039242.

[57] Andy D. Perkins and Michael A. Langston. Threshold selection in gene co-expression networks using spectral graph theory techniques. *BMC Bioinformatics*, 10(11):S4, Oct 2009. ISSN 1471-2105. doi: 10.1186/1471-2105-10-S11-S4.

[58] Belinda Phipson and Gordon K. Smyth. Permutation P-values Should Never Be Zero: Calculating Exact P-values When Permutations Are Randomly Drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1), January 2010. doi: 10.2202/1544-6115.1585. URL https://doi.org/10.2202/1544-6115.1585.

[59] Nicolle Rager. How proteins are made NSF. National Science Foundation, Public Domain, 2012. URL https://commons.wikimedia.org/wiki/File:How_proteins_are_made_NSF.jpg. File: How proteins are made NSF.jpg.

[60] C. J. Robinson, B. J. M. Bohannan, and V. B. Young. From Structure to Function: the Ecology of Host-Associated Microbial Communities. *Microbiology and Molecular Biology Reviews*, 74(3):453–476, August 2010. doi: 10.1128/mmbr.00014-10. URL https://doi.org/10.1128/mmbr.00014-10.

[61] Francisco J. Romero-Campero, Ignacio Perez-Hurtado, Eva Lucas-Reina, Jose M. Romero, and Federico Valverde. ChlamyNET: a Chlamydomonas gene co-expression network reveals global properties of the transcriptome and the early setup of key co-expression patterns in the green lineage. *BMC Genomics*, 17(1), March 2016. doi: 10.1186/s12864-016-2564-y. URL https://doi.org/10.1186/s12864-016-2564-y.

[62] Matthew C. Ross, Donna M. Muzny, Joseph B. McCormick, Richard A. Gibbs, Susan P. Fisher-Hoch, and Joseph F. Petrosino. 16s gut community of the Cameron County Hispanic Cohort. *Microbiome*, 3(1):7, 2015. doi: 10.1186/s40168-015-0072-y. URL https://doi.org/10.1186/s40168-015-0072-y.

[63] W. Matthew Sattley and Michael T. Madigan. *Microbiology*, pages 1–10. American Cancer Society, 2015. ISBN 9780470015902. doi: 10.1002/9780470015902.a0000459.pub2. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470015902.a0000459.pub2.

[64] Andrea Sboner, Xinmeng Jasmine Mu, Dov Greenbaum, Raymond K. Auerbach, and Mark B. Gerstein. The real cost of sequencing: higher than you think! *Genome Biology*, 12 (8):125, Aug 2011. ISSN 1474-760X. doi: 10.1186/gb-2011-12-8-125. URL https://doi.org/10.1186/gb-2011-12-8-125.

[65] Filip Scheperjans, Velma Aho, Pedro A. B. Pereira, Kaisa Koskinen, Lars Paulin, Eero Pekkonen, Elena Haapaniemi, Seppo Kaakkola, Johanna Eerola-Rautio, Marjatta Pohja, Esko Kinnunen, Kari Murros, and Petri Auvinen. Gut microbiota are related to Parkinson's disease and clinical phenotype. *Movement Disorders*, 30(3):350–358, December 2014. doi: 10.1002/mds.26069. URL https://doi.org/10.1002/mds.26069.

[66] Jose U. Scher, Andrew Sczesnak, Randy S. Longman, Nicola Segata, Carles Ubeda, Craig Bielski, Tim Rostron, Vincenzo Cerundolo, Eric G. Pamer, Steven B. Abramson, Curtis Huttenhower, and Dan R. Littman. Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis. *eLife*, 2, November 2013. doi: 10.7554/elife.01202. URL https://doi.org/10.7554/elife.01202.

[67] Alyxandria M. Schubert, Mary A. M. Rogers, Cathrin Ring, Jill Mogle, Joseph P. Petrosino, Vincent B. Young, David M. Aronoff, and Patrick D. Schloss. Microbiome Data Distinguish Patients with Clostridium difficile Infection and Non-C. difficile-Associated Diarrhea from Healthy Controls. *mBio*, 5(3), May 2014. doi: 10.1128/mbio.01021-14. URL https://doi.org/10.1128/mbio.01021-14.

[68] Ron Sender, Shai Fuchs, and Ron Milo. Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans. *Cell*, 164(3):337–340, January 2016. doi: 10.1016/j.cell.2016.01.013. URL https://doi.org/10.1016/j.cell.2016.01.013.

[69] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, 13(11):2498–2504, November 2003. doi: 10.1101/gr.1239303. URL https://doi.org/10.1101/gr.1239303.

[70] Pallavi Singh, Tracy K. Teal, Terence L. Marsh, James M. Tiedje, Rebekah Mosci, Katherine Jernigan, Angela Zell, Duane W. Newton, Hossein Salimnia, Paul Lephart, Daniel Sundin, Walid Khalife, Robert A. Britton, James T. Rudrik, and Shannon D. Manning. Intestinal microbial communities associated with acute enteric infections and disease recovery. *Microbiome*, 3(1):45, Sep 2015. ISSN 2049-2618. doi: 10.1186/s40168-015-0109-2. URL https://doi.org/10.1186/s40168-015-0109-2.

[71] Joshua S. Son, Ling J. Zheng, Leahana M. Rowehl, Xinyu Tian, Yuanhao Zhang, Wei Zhu, Leighann Litcher-Kelly, Kenneth D. Gadow, Grace Gathungu, Charles E. Robertson, Diana Ir, Daniel N. Frank, and Ellen Li. Comparison of Fecal Microbiota in Children with Autism Spectrum Disorders and Neurotypical Siblings in the Simons Simplex Collection. *PLOS ONE*, 10(10):e0137725, October 2015. doi: 10.1371/journal.pone.0137725. URL https://doi.org/10.1371/journal.pone.0137725.

[72] Fei Teng, Sree Sankar Darveekaran Nair, Pengfei Zhu, Shanshan Li, Shi Huang, Xiaolan Li, Jian Xu, and Fang Yang. Impact of DNA extraction method and targeted 16S-rRNA hypervariable region on oral microbiota profiling. *Scientific Reports*, 8(1), November 2018. doi: 10.1038/s41598-018-34294-x. URL https://doi.org/10.1038/s41598-018-34294-x.

[73] Peter J. Turnbaugh, Micah Hamady, Tanya Yatsunenko, Brandi L. Cantarel, Alexis Duncan, Ruth E. Ley, Mitchell L. Sogin, William J. Jones, Bruce A. Roe, Jason P. Affourtit, Michael Egholm, Bernard Henrissat, Andrew C. Heath, Rob Knight, and Jeffrey I. Gordon. A core gut microbiome in obese and lean twins. *Nature*, 457(7228):480–484, November 2008. doi: 10.1038/nature07540. URL https://doi.org/10.1038/nature07540.

[74] Alessandro Vespignani. Twenty years of network science. *Nature*, 558, 06 2018. doi: 10.1038/d41586-018-05444-y.

[75] Marc Vidal. Human Cell. Own work, CC BY-SA 4.0 License, 2013. URL https://commons.wikimedia.org/wiki/File:Human_Cell.jpg. File: Human Cell.jpg.

[76] Caroline Vincent, David A. Stephens, Vivian G. Loo, Thaddeus J. Edens, Marcel A. Behr, Ken Dewar, and Amee R. Manges. Reductions in intestinal Clostridiales precede the development of nosocomial Clostridium difficile infection. *Microbiome*, 1(1):18, Jun 2013. ISSN 2049-2618. doi: 10.1186/2049-2618-1-18. URL https://doi.org/10.1186/2049-2618-1-18.

[77] Qiong Wang, George M. Garrity, James M. Tiedje, and James R. Cole. Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Applied and Environmental Microbiology*, 73(16):5261–5267, 2007. ISSN 0099-2240. doi: 10.1128/AEM.00062-07. URL https://aem.asm.org/content/73/16/5261.

[78] Tingting Wang, Guoxiang Cai, Yunping Qiu, Na Fei, Menghui Zhang, Xiaoyan Pang, Wei Jia, Sanjun Cai, and Liping Zhao. Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. *The ISME Journal*, 6(2):320–329, August 2011. doi: 10.1038/ismej.2011.109. URL https://doi.org/10.1038/ismej.2011.109.

[79] Ying Wang, Haiyan Hu, and Xiaoman Li. rRNAFilter: A Fast Approach for Ribosomal RNA Read Removal Without a Reference Database. *Journal of Computational Biology*, 24 (4):368–375, April 2017. doi: 10.1089/cmb.2016.0113. URL https://doi.org/10.1089/cmb.2016.0113.

[80] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, June 1998. doi: 10.1038/30918. URL https://doi.org/10.1038/30918.

[81] Stephen C Watts, Kathryn E Holt, Michael Inouye, and Scott C Ritchie. FastSpar: rapid and scalable correlation estimation for compositional data. *Bioinformatics*, 35(6):1064–1066, 08 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty734. URL https://doi.org/10.1093/bioinformatics/bty734.

[82] Sophie Weiss, Zhenjiang Zech Xu, Shyamal Peddada, Amnon Amir, Kyle Bittinger, Antonio Gonzalez, Catherine Lozupone, Jesse R. Zaneveld, Yoshiki Vázquez-Baeza, Amanda Birmingham, Embriette R. Hyde, and Rob Knight. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, 5(1), March 2017. doi: 10.1186/s40168-017-0237-y. URL https://doi.org/10.1186/s40168-017-0237-y.

[83] Brian T. Werner. Complexity in Natural Landform Patterns. *Science*, 284(5411):102–104, April 1999. doi: 10.1126/science.284.5411.102. URL https://doi.org/10.1126/science.284.5411.102.

[84] Ben P. Willing, Johan Dicksved, Jonas Halfvarson, Anders F. Andersson, Marianna Lucio, Zongli Zheng, Gunnar Järnerot, Curt Tysk, Janet K. Jansson, and Lars Engstrand. A Pyrosequencing Study in Twins Shows That Gastrointestinal Microbial Profiles Vary With Inflammatory Bowel Disease Phenotypes. *Gastroenterology*, 139(6):1844–1854.e1, December 2010. doi: 10.1053/j.gastro.2010.08.049. URL https://doi.org/10.1053/j.gastro.2010.08.049.

[85] Vincent Wai-Sun Wong, Chi-Hang Tse, Tommy Tsan-Yuk Lam, Grace Lai-Hung Wong, Angel Mei-Ling Chim, Winnie Chiu-Wing Chu, David Ka-Wai Yeung, Patrick Tik-Wan Law, Hoi-Shan Kwan, Jun Yu, Joseph Jao-Yiu Sung, and Henry Lik-Yuen Chan. Molecular Characterization of the Fecal Microbiota in Patients with Nonalcoholic Steatohepatitis – A Longitudinal Study. *PLoS ONE*, 8(4):e62885, April 2013. doi: 10.1371/journal.pone.0062885. URL https://doi.org/10.1371/journal.pone.0062885.

[86] Ilan Youngster, Jenny Sauk, Christina Pindar, Robin G. Wilson, Jess L. Kaplan, Mark B. Smith, Eric J. Alm, Dirk Gevers, George H. Russell, and Elizabeth L. Hohmann. Fecal Microbiota Transplant for Relapsing Clostridium difficile Infection Using a Frozen Inoculum From Unrelated Donors: A Randomized, Open-Label, Controlled Pilot Study. *Clinical Infectious Diseases*, 58(11):1515–1522, April 2014. doi: 10.1093/cid/ciu135. URL https://doi.org/10.1093/cid/ciu135.

[87] G. Zeller, J. Tap, A. Y. Voigt, S. Sunagawa, J. R. Kultima, P. I. Costea, A. Amiot, J. Bohm, F. Brunetti, N. Habermann, R. Hercog, M. Koch, A. Luciani, D. R. Mende, M. A. Schneider, P. Schrotz-King, C. Tournigand, J. Tran Van Nhieu, T. Yamada, J. Zimmermann, V. Benes, M. Kloor, C. M. Ulrich, M. von Knebel Doeberitz, I. Sobhani, and P. Bork. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular Systems Biology*, 10(11):766–766, November 2014. doi: 10.15252/msb.20145645. URL https://doi.org/10.15252/msb.20145645.

[88] Zhigang Zhang, Huiqin Zhai, Jiawei Geng, Rui Yu, Haiqing Ren, Hong Fan, and Peng Shi. Large-Scale Survey of Gut Microbiota Associated With MHE Via 16S rRNA-Based Pyrosequencing. *The American Journal of Gastroenterology*, 108(10):1601–1611, July 2013. doi: 10.1038/ajg.2013.221. URL https://doi.org/10.1038/ajg.2013.221.

[89] Jizhong Zhou, Wenzong Liu, Ye Deng, Yi-Huei Jiang, Kai Xue, Zhili He, Joy D. Van Nostrand, Liyou Wu, Yunfeng Yang, and Aijie Wang. Stochastic Assembly Leads to Alternative Communities with Distinct Functions in a Bioreactor Microbial Community. *mBio*, 4(2), March 2013. doi: 10.1128/mbio.00584-12. URL https://doi.org/10.1128/mbio.00584-12.

[90] Lixin Zhu, Susan S. Baker, Chelsea Gill, Wensheng Liu, Razan Alkhouri, Robert D. Baker, and Steven R. Gill. Characterization of gut microbiomes in nonalcoholic steatohepatitis (NASH) patients: A connection between endogenous alcohol and NASH. *Hepatology*, 57(2): 601–609, January 2013. doi: 10.1002/hep.26093. URL https://doi.org/10.1002/hep.26093.

[91] Margaret L. Zupancic, Brandi L. Cantarel, Zhenqiu Liu, Elliott F. Drabek, Kathleen A. Ryan, Shana Cirimotich, Cheron Jones, Rob Knight, William A. Walters, Daniel Knights, Emmanuel F. Mongodin, Richard B. Horenstein, Braxton D. Mitchell, Nanette Steinle, Soren Snitker, Alan R. Shuldiner, and Claire M. Fraser. Analysis of the Gut Microbiota in the Old Order Amish and Its Relation to the Metabolic Syndrome. *PLoS ONE*, 7(8):e43052, August 2012. doi: 10.1371/journal.pone.0043052. URL https://doi.org/10.1371/journal.pone.0043052.

A

Supplementary Figures

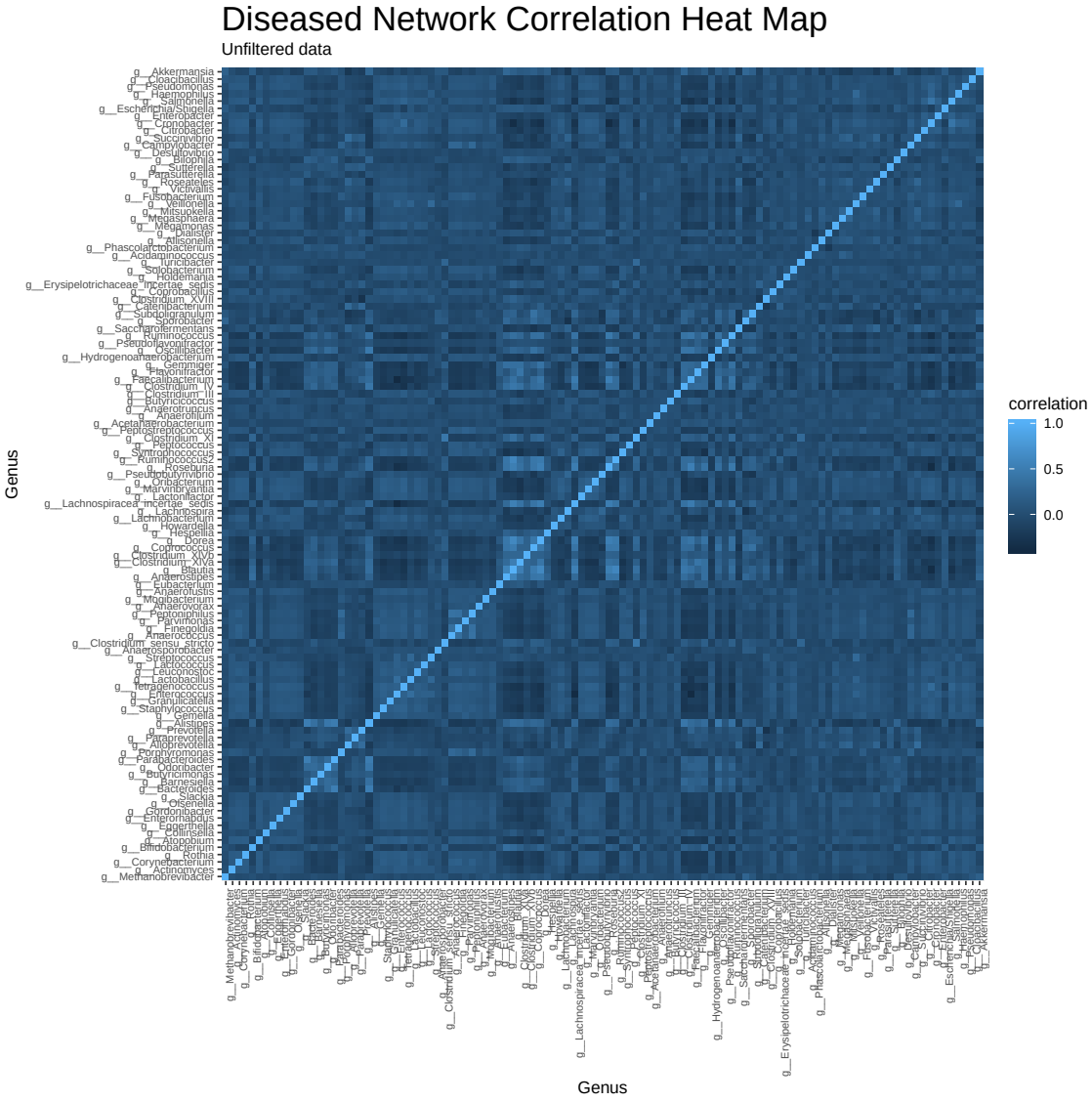
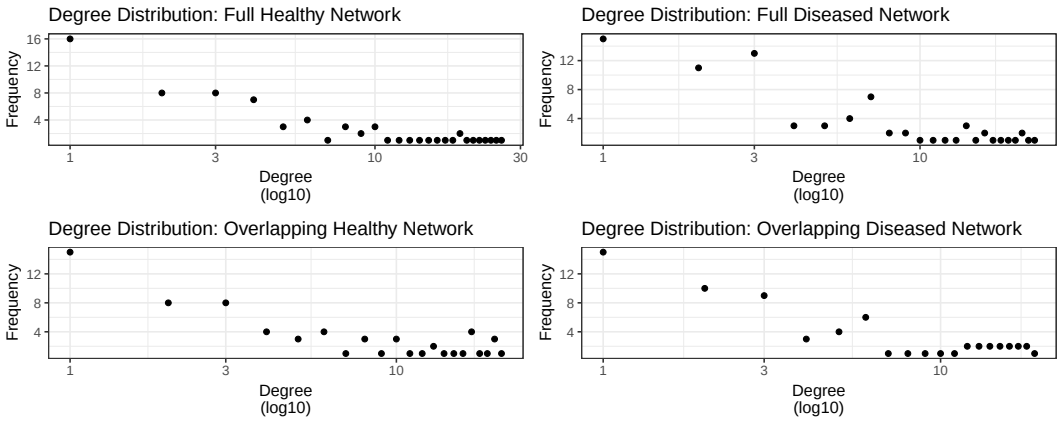
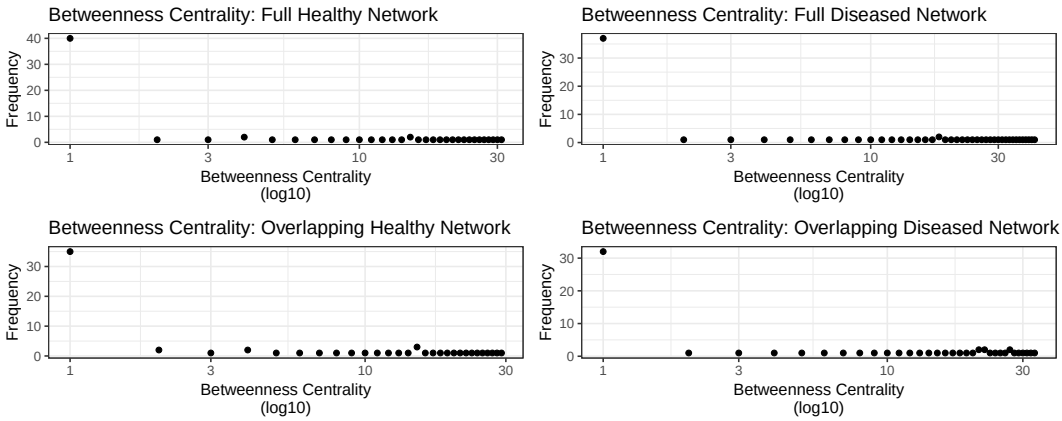


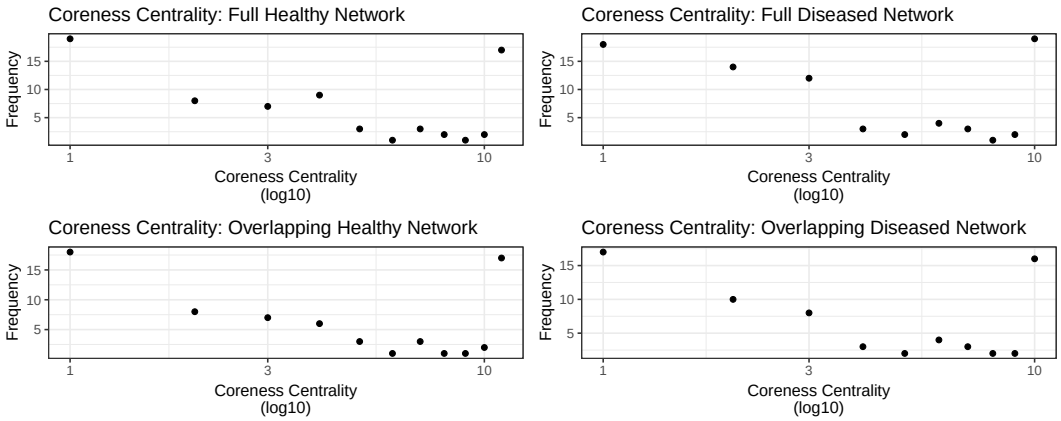
Figure A.1: Heat map of the resulting diseased FastSpar correlation matrix. Genera are listed on the respective axes and correlation values are colored based upon their strength.



(a) Degree distribution for the filtered healthy and diseased networks and the overlapping sub-graphs.



(b) Betweenness centrality distribution for the filtered healthy and diseased networks and the overlapping sub-graphs



(c) Coreness centrality distribution for the filtered healthy and diseased networks and the overlapping sub-graphs

Figure A.2: Distributions mentioned in Section 5.3 for the filtered healthy and diseased networks. **(a)** contains the degree distributions, **(b)** contains the betweenness centrality distributions, and **(c)** contains the coreness centrality distributions.

	Degree FHN	Degree FDN
Rank 1	38, g_Clostridium_IV	40, g_Clostridium_IV
Rank 2	33, g_Clostridium_XlVa	30, g_Coprococcus
Rank 3	32, g_Blautia	29, g_Clostridium_XlVa
Rank 4	31, g_Coprococcus	29, g_Ruminococcus
Rank 5	29, g_Gemmiger	28, g_Blautia
	Degree OHN	Degree ODN
Rank 1	35, g_Clostridium_IV	35, g_Clostridium_IV
Rank 2	29, g_Blautia	27, g_Coprococcus
Rank 3	29, g_Clostridium_XlVa	27, g_Ruminococcus
Rank 4	29, g_Coprococcus	26, g_Blautia
Rank 5	27, g_Ruminococcus	26, g_Clostridium_XlVa
	Betweenness Centrality FHN	Betweenness Centrality FDN
Rank 1	504, g_Clostridium_IV	626, g_Clostridium_IV
Rank 2	489, g_Blautia	577, g_Coprococcus
Rank 3	319, g_Clostridium_XlVa	344, g_Clostridium_XlVa
Rank 4	308, g_Oscillibacter	324, g_Blautia
Rank 5	254, g_Faecalibacterium	250, g_Faecalibacterium
	Betweenness Centrality OHN	Betweenness Centrality ODN
Rank 1	428, g_Blautia	438, g_Clostridium_IV
Rank 2	388, g_Clostridium_IV	297, g_Blautia
Rank 3	294, g_Oscillibacter	222, g_Clostridium_XlVa
Rank 4	249, g_Clostridium_XlVa	205, g_Faecalibacterium
Rank 5	221, g_Sporobacter	203, g_Coprococcus
	Coreness Centrality FHN	Coreness Centrality FDN
Rank 1	12, g_Alistipes	11, g_Alistipes
Rank 2	12, g_Anaerostipes	11, g_Tetragenococcus
Rank 3	12, g_Blautia	11, g_Anaerostipes
Rank 4	12, g_Clostridium_XlVa	11, g_Blautia
Rank 5	12, g_Coprococcus	11, g_Clostridium_XlVa
	Coreness Centrality OHN	Coreness Centrality ODN
Rank 1	12, g_Alistipes	11, g_Alistipes
Rank 2	12, g_Anaerostipes	11, g_Anaerostipes
Rank 3	12, g_Blautia	11, g_Blautia
Rank 4	12, g_Clostridium_XlVa	11, g_Clostridium_XlVa
Rank 5	12, g_Coprococcus	11, g_Coprococcus

Table A.1: Table mentioned in Section 5.3 listing the five genera that have the highest degree, betweenness centrality, and coreness centrality for the full healthy network (FHN), full diseased network (FDN), overlapping healthy network (OHN), and overlapping diseased network (ODN).

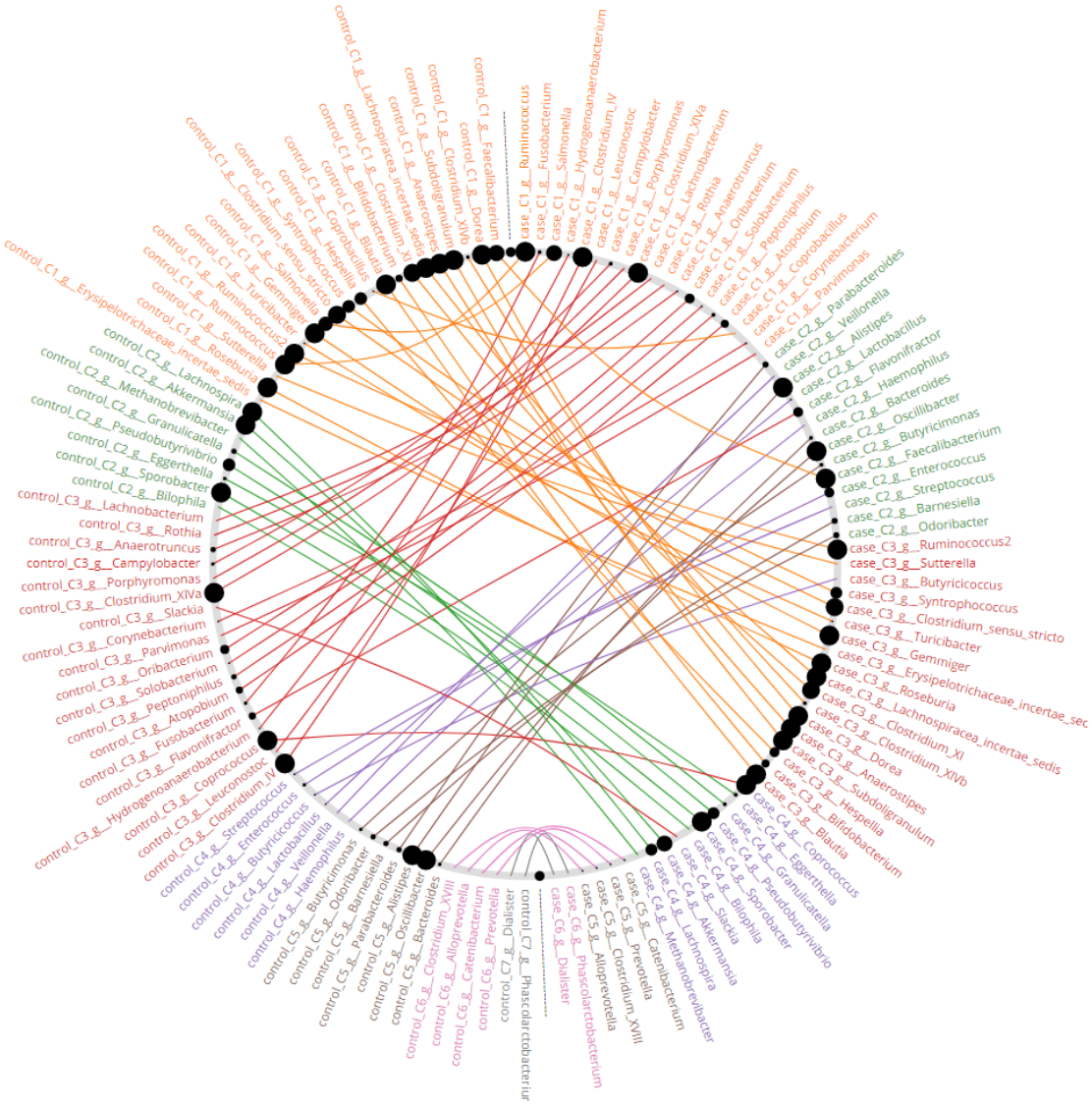


Figure A.3: The full NetShift shuffle diagram highlighting re-wiring between communities identified by hierarchical clustering mentioned in Section 5.4. All taxa are represented by the control taxa on the left of the diagram and case taxa on the right. Communities are identified by the color of the taxon. The shuffle plot in Figure 5.7 can be useful to interpret this diagram.