Human interpretable artificial intelligence applications for microbial-related diseases

by

Josh L. Espinoza, M.Sc

Submitted in fulfilment of the requirements of the degree of Doctor of Philosophy of Applied Science in Biotechnology in the Faculty of Applied Sciences at the Durban University of Technology

Supervisor: Suren Singh, PhD (*Durban University of Technology)*
Co-supervisor: Karen E. Nelson (*J. Craig Venter Institute*)
Co-supervisor: Chris L. Dupont (*J. Craig Venter Institute*)

2022

DECLARATION

I, **Josh L. Espinoza**, declare that the thesis submitted for the degree of Doctor of Philosophy in Biotechnology at the Durban University of Technology is the result of my investigation and has not already been accepted in substance for any degree and is not being concurrently submitted for any other degree. All the work was done by the author.

Student: Josh L. Espinoza

Student Signature:

Supervisor: Suren Singh

Supervisor Signature:

Co-Supervisor: Karen E. Nelson

Co-Supervisor Signature:

Co-Supervisor: Christopher L. Dupont

Co-Supervisor Signature:

ABSTRACT

The human microbiome is a complex ecosystem that is influenced not only by host genetics but environmental stimuli. With advancements in next-generation sequencing (NGS) technologies, genomics and related meta-omics such as metagenomics, metatranscriptomics and metaproteomics have become increasingly accessible for researchers and clinicians to investigate microbial-related diseases. However, analysis of the outputs of “omics” technologies are often difficult due to variance introduced by biological complexity, batch effects from laboratory protocols/conditions, and the sensitivity/calibration of highly sensitive instruments. The biological complexity of “omics” presents a considerable analytical obstacle as most datasets contain hundreds of thousands to millions of unique features with unknown connections and nested hierarchies. In addition to this inherent complexity, the deluge of data generated from NGS technologies is fundamentally compositional, conveys only relative information, and because of this cannot be robustly analyzed using conventional statistical approaches. Furthermore, meta-omics datasets are typically sparse and the number of biological features often vastly exceeds the number of biological samples which can introduce anomalies in statistical analysis and the downstream findings if not addressed accordingly; a term dubbed as “the curse of dimensionality”. The complexity, compositionality, and dimensionality of “omics” datasets makes it challenging to derive clinical meaning and an understanding of the microbial system with respect to a host phenotype.

Although, artificial intelligence and machine-learning methods have progressed substantially in recent years, their applications in domain sciences such as biology, and by extension “omics” technologies, have been limited in terms of human interpretability. In many machine-learning paradigms, interpretability is often sacrificed for analytical performance, or vice versa, but recently a domain-agnostic effort aims to develop explainable artificial intelligence algorithms that have both high modeling performance and human interpretability; a major goal of biomedical sciences.

In this dissertation, I develop novel approaches in bridging biological science with machine learning methods at the vanguard of scientific development through the initiative of explainable artificial intelligence. The methods developed are validated on 3 datasets pertaining to microbial-related diseases including antibiotic resistance discovery, acute malnutrition in West African children, and caries pathology in Australian juvenile twins. The combination of methods developed are expected to provide the means for clinical researchers to overcome obstacles in interrogating the complex narratives that determine health and disease.

DEDICATION

Foremost, I dedicate this research to my family, my friends, and my colleagues for all their support, wisdom, and inspiration during this work. In particular, my partner (and now wife) Madison Espinoza (formerly Teague) who has been along for the ride during this bioinformatics journey. Pursuing a doctorate degree, working full-time, and planning a wedding during a “once in 100-year pandemic” was not the easiest feat to accomplish and I could not have imagined doing this without her support. I would also like to dedicate this work to my mother for always encouraging me to pursue my passions in science. My mother has been an essential source of guidance ever since I began crafting my professional career prompting me to consider where each path would lead while also giving unconditional support.

Further, I dedicate this work to all the sacrifices made by those who have made it their personal mission to leave this world better than how they found it. Whether through scientific advancements, environmental preservation, humanitarian efforts, or interpersonal relationships, we all possess the ability to make a positive and lasting impact.

Lastly, I would like to dedicate this research to the natural world that continues to inspire me. Whenever I am impeded by a mental barrier or a lack of motivation, my intrigue and respect for the natural world reminds me of why I must persist and persevere to address the most pressing ecological and humanitarian issues.

ACKNOWLEDGEMENTS

My research journey through joint appointments at Durban University of Technology and the J. Craig Venter Institute has been extremely rewarding. Although my dissertation was written solo, this collection of work involved collaboration with various domain experts, wisdom from colleagues, and countless hours dedicated by my supervisors to secure funding. There are many individuals ranging from classmates to collaborators who have played significant roles in my research journey but I would like to acknowledge the most prominent.

First, I would like to have the honor to acknowledge my current advisors at the Durban University of Technology and J. Craig Venter Institute for their mentorship throughout these projects. Chris L. Dupont, Karen E. Nelson, and Suren Singh have challenged and guided me through a multitude of projects each with their own nuances and opportunities to develop new techniques. While providing me with a solid biological perspective, my advisors have also given me the freedom to explore complex problems in creative ways while working within the bounds of statistical assumptions with respect to each dataset. In particular, I have worked with Chris since I was an intern - before receiving my Master's degree - and our synergy to conduct research on a diverse array of topics ranging from vaccine development to coral-related climate change to characterizing the human microbiome has been and continues to be an enlightening path into the future. His balance of patience and progress has been essential in my research endeavors. Karen has also been a key component in my research career and continues as a role model in pushing the boundaries of science through her accomplishments. Suren Singh has

provided me with outstanding mentorship, several research opportunities, and unique perspectives from a biotechnology point of view.

Second, I would like to acknowledge my previous stem cell research advisors Eleen Y. Shum, Miles F. Wilkinson, and Chih H. Lou (Robert) from the University of California, San Diego who have provided me with a solid foundation for experimental design, an outlet to explore my interest in bioinformatics, and the critical nudge in shifting from molecular biology to bioinformatics. Miles was the first PI I had that taught me how to “think like a scientist” in the types of hypotheses I would develop and how I would test these hypotheses. Robert was my first direct mentors and was an essential catalyst in my transition to bioinformatics as we explored his hypotheses *in silico*. Eleen was my unofficial advisor towards the latter half of my time in the Wilkinson lab and she devoted much of her time during her own PhD teaching me the utility of statistical modeling, incorporated me into several research projects, and has been a beacon of support both from a career perspective and as a friend. In this vein, Susan Kaiser who was the program administrator for the Stem Cell Internship Program at SDSU was instrumental in my scientific career.

Lastly, I would like to acknowledge my advisors Elizabeth Dinsdale and Scott Kelley from San Diego State University - during my undergraduate studies - for providing the opportunity to evolve my nascent programming skills towards my passion in marine ecology. In 2019, I had the pleasure of speaking at an international conference in the timeslot directly after Liz which was an extremely rewarding experience given that she

was one of my earliest mentors and introduced me to the marvel of metagenomics and marine science.

TABLE OF CONTENTS

PEER-REVIEWED PUBLICATIONS

This dissertation is based on the following peer-reviewed publications that are referred to throughout the text by their Roman numerical delineation.

+ indicates equal contribution

Primary Publications:

I. **Espinoza JL**, Shah N, Singh S, Nelson KE, Dupont CL. Applications of weighted association networks applied to compositional data in biology. Environ Microbiol. 2020 May 20;. doi: 10.1111/1462-2920.15091. Review. PubMed PMID: 32436334.

II. **Espinoza JL+**, Dupont CL+, O'Rourke A, Beyhan S, Morales P, Spoering A, Meyer KJ, Chan AP, Choi Y, Nierman WC, Lewis K, Nelson KE. Predicting antimicrobial mechanism-of-action from transcriptomes: A generalizable explainable artificial intelligence approach. PLoS Comput Biol. 2021 Mar 29;17(3):e1008857. doi: 10.1371/journal.pcbi.1008857. PubMed PMID: 33780444.

III. Nabwera HM+, **Espinoza JL+**, Worwui A, Betts M, Okoi C, Sesay AK, Bancroft R, Agbla SC, Jarju S, Bradbury RS, Colley M, Jallow AT, Liu J, Houpt ER, Prentice AM, Antonio M, Bernstein RM, Dupont CL+, Kwambana-Adams BA+. Interactions between fecal gut microbiome, enteric pathogens, and energy regulating hormones among acutely malnourished rural Gambian children. EBioMedicine. 2021 Oct 22;73:103644. doi: 10.1016/j.ebiom.2021.103644. PMID: 34695658.

IV. **Espinoza JL**, Torralba MG, Leong P, Saffery R, Bockmann M, Kuelbs C, Singh S, Hughes T, Craig JM, Nelson KE, Dupont CL. Differential network analysis of oral microbiome metatranscriptomes identifies community scale metabolic restructuring in dental caries. Submitted to PNAS Nexus.

Secondary Publications:

V. Santoro EP, Borges RM, **Espinoza JL**, Freire M, Messias CSMA, Villela HDM, Pereira LM, Vilela CLS, Rosado JG, Cardoso PM, Rosado PM, Assis JM, Duarte GAS, Perna G, Rosado AS, Macrae A, Dupont CL, Nelson KE, Sweet MJ, Voolstra CR, Peixoto RS. Coral microbiome manipulation elicits metabolic and genetic restructuring to mitigate heat stress and evade mortality. Sci Adv. 2021 Aug 13;7(33):eabg3088. doi: 10.1126/sciadv.abg3088. PMID: 34389536.

INTERNATIONAL CONFERENCES AND SEMINARS

- **Espinoza JL**, Torralba MG, Leong P, Saffery R, Bockmann M, Kuelbs C, Singh S, Hughes T, Craig JM, Nelson KE, Dupont CL. Microbiome feature engineering and comparative network analysis. *Oral Presentation at EMBO – The Human Microbiome Conference*. National Institute of Biomedical Genomics. Kalyani, India. 13 November 2019
- **Espinoza JL**, Dupont CL, O'Rourke A, Beyhan S, Morales P, Spoering A, Meyer KJ, Chan AP, Choi Y, Nierman WC, Lewis K, Nelson KE. Predicting antimicrobial mechanism-of-action from transcriptomes: A generalizable explainable artificial intelligence approach. *Oral Presentation at EBRC – Virtual Seminar Series*. Virtual per COVID-19 regulations. 18 August 2020

- **Espinoza JL**, Dupont CL, O'Rourke A, Beyhan S, Morales P, Spoering A, Meyer KJ, Chan AP, Choi Y, Nierman WC, Lewis K, Nelson KE. Predicting antimicrobial mechanism-of-action from transcriptomes: A generalizable explainable artificial intelligence approach. *Oral Presentation at NASA – Lunch, Learn, and Discuss Seminar Series*. Virtual per COVID-19 regulations. 20 July 2020
- **Espinoza JL**, Shah N, Nelson KE, Dupont CL. Leveraging feature selection algorithms for phenotype-specific community detection. *Oral Presentation at JCVI WIP Seminar Series*. J. Craig Venter Institute. La Jolla, CA USA. 21 November 2019
- **Espinoza JL**, Nabwera HM, Worwui A, Betts M, Okoi C, Sesay AK, Bancroft R, Agbla SC, Jarju S, Bradbury RS, Colley M, Jallow AT, Liu J, Houpt ER, Prentice AM, Antonio M, Bernstein RM, Dupont CL, Kwambana-Adams BA. Interactions between fecal gut microbiome, enteric pathogens, and energy regulating hormones among acutely malnourished rural Gambian children. *Poster Presentation at VKS – Harness the Microbiome for Disease Prevention and Therapy*. Virtual per COVID-19 regulations. 19 January 2020

LIST OF ABBREVIATIONS AND ACRYONYMS

(Co)HEC	(Clairvoyance-optimized) Hierarchical Ensemble of Classifiers
(X)AI	(Explainable) Artificial Intelligence
Δ	Difference between two conditions
ALR	Additive Log-Ratio
AN	Aggregate Networks
ANI	Average Nucleotide Identity
ASV	Amplicon Sequence Variant
CDC	Center for Disease Control
CLR	Center Log-Ratio
CoDA	Compositional Data Analysis
CPR	Candidate Phyla Radiation
DARPA	Defense Advanced Research Projects Agency
DCN	Differential Coexpression Network
DGE	Differential Gene Expression
FDR	False Discovery Rate
HCDC	High Connectivity DCN Cluster
ILR	Isometric Log-Ratio
k	Connectivity
LCOCV	Leave Compound Out Cross Validation
logFC	Log Fold Change
MAG	Metagenome assembled genome
MAM	Moderately Acute Malnourished
MOA	Mechanism of Action
MSE	Mean Squared Error
NGS	Next-Generation Sequencing
NIH	National Institute of Health
OTU	Operational Taxonomic Unit
PGFC	Phylogenomic Functional Category
PSCN	Phenotype Specific Coexpression Network
rRNA	Ribosomal RNA
SAM	Severely Acute Malnourished
SEM	Standard Error of the Mean
SLC	Species Level Cluster
SSPN	Sample Specific Perturbation Network
WHO	World Health Organization
WHZ	Weight-for-height Z-score
WN	Well-Nourished
μ	Mean of distribution

LIST OF FIGURES

LIST OF TABLES

1 - INTRODUCTION

1.1 – Background and Context

Microbial-related diseases are a critical concern for public health. The study of these diseases is often stymied by a battery of obstacles ranging from high-level interpretation of microbial complexes identified *in silico* to the low-level assumptions of data generated by next-generation sequencing (NGS) technologies. While each microbial-related disease is unique in its own, the methodologies used to investigate these diseases are not disease specific. NGS instruments produce a wealth of data and can be analyzed *in silico* to provide insight into not only what it means biologically to be diagnosed with a disease but also what it means to be healthy in the context of said disease. However, this wealth is not without its caveats as this “embarrassment of riches” often leads to a dimensionality where the number of features vastly exceeds the number of observations. Analysis that does not address this dimensionality obstacle may result in unreliable findings that are not generalizable to other similar datasets.

Traditional approaches often focus solely on the abundance of individual features (e.g., genes, microbes, etc.) when studying microbial-related diseases. Although this can reveal biological insight and is suitable to address many hypotheses, abundance-centric approaches often fail to provide insight into the microbial community structure and how this topology changes between healthy and unhealthy phenotypes, many of which include a multitude of microbial configurations that are grouped within the context of an umbrella diagnosis. Investigating microbial communities through the confines of compositional data analysis (CoDA), a critical framework when investigating NGS data especially the

context of network analysis, is an area of active albeit nascent research and growing amongst microbial ecologists with the advent of software packages built on a solid foundation of theoretical mathematics. The advantages of network theory in the context of microbial ecology are vast but the interpretations from a biological or clinical perspective are often precarious as the number of pairwise interactions amongst features increases quadratically with respect to the number of features.

Machine learning, a subset of artificial intelligence (AI), is an invaluable resource for distilling biological meaning from large and complex datasets such as NGS datasets and/or their network representations. However, these methodologies are by no means a silver bullet; failure to address assumptions of NGS data have a high likelihood of yielding spurious conclusions. In particular, the goal of many NGS experiments is to detect as many high-quality biological features as possible and, despite a decline in experimental costs, this facet of the experimental design often leads to a far greater number of features than observations, thus, increasing the noise to signal ratio; a property exacerbated when the category of observations is imbalanced which is the norm in most real-world studies. Proper implementations of machine learning approaches to study such datasets should include methodologies that address this dimensionality obstacle to minimize the risk of spurious results. Further, many existing approaches do not recognize the compositionality of NGS data and therefore do not abide by the assumptions imposed by CoDA fundamentals. Therefore, findings based on statistical fallacies that do not address dimensionality obstacles and compositionality may be misleading when it comes to biological or clinical interpretation.

The performance of machine learning algorithms increases with algorithm complexity making interpretation of AI methodologies non-trivial. Explainable AI (XAI) is a paradigm in which algorithms are designed for both high performance and human interpretability. These algorithms function as a feed-forward knowledge loop where better understanding of domain expertise (e.g., biology, chemistry) result in more realistic models which, in turn, lead to a better understanding of the underlying process. XAI as a synergy of domain expertise and machine learning presents a productive framework to characterize microbial-related diseases by leveraging the wealth of data generated from NGS instruments while also incorporating the many assumptions that are inherent in NGS datasets.

1.2 – Aim and Objectives

Aim:

To develop explainable artificial intelligence methodologies that can be used to characterize and provide insight into the mechanisms of microbial-related diseases.

Objectives:

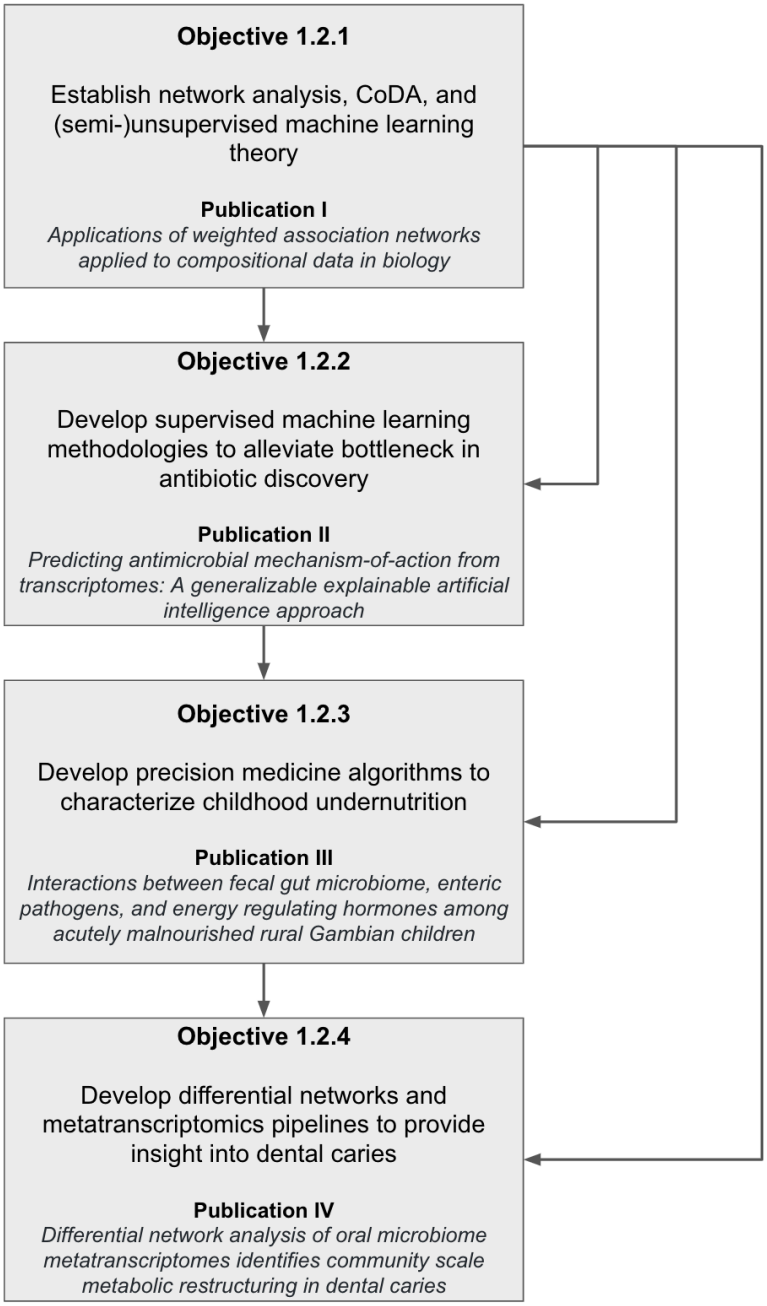
1.2.1 **–** Compile existing literature to establish a theoretical foundation for compositional data analysis, network theory, and (semi-)unsupervised machine learning for investigating next-generation sequencing for biotechnology. Accomplished in Publication I (*Environmental Microbiology*).

1.2.2 **–** Development of explainable supervised machine learning algorithms to broadly address the dimensionality obstacles inherently in next-generation sequencing datasets while applying these methodologies to alleviate the bottleneck in antibiotic discovery pipelines. Accomplished in Publication II (*PLOS Computational Biology*).

1.2.3 **–** Development of explainable precision medicine algorithms and demonstrate their utility in characterizing childhood undernutrition using clinical measurements and fecal microbiome sequencing. Accomplished in Publication III (*EBioMedicine*).

1.2.4 **–** Development of differential network analysis algorithms and scalable metatranscriptomics pipelines while using these approaches synergistically to characterize the oral microbiome in relation to dental caries. Accomplished in Manuscript IV pending publication.

1.3 – Dissertation Structure



2 - LITERATURE REVIEW

2.1 - From organisms to ecosystems

Life can be visualized as a hierarchy as one transitions from molecules to unicellular organisms and from individual organisms to ecosystems. An ecosystem is a complex of organisms interacting with one another and with the physical environment that defines the ecosystem with stability governed by a system of checks and balances. Ecosystems are not constrained to environmental realms and can be defined over a broader domain especially in the context of microbiology. For instance, the human microbiome is a dynamic ecosystem comprised of microorganisms that have colonized the human body and many of which interact with host human metabolism either directly or indirectly. As with any environment, the human microbiome can be partitioned into more distinct macroenvironment ecosystems such as the oral, nasopharyngeal, gut, vaginal, and skin microbiomes each with their own microenvironments (e.g., supragingival vs. subgingival plaque microbiomes) (Fig. 1).

A common myth, and fallacy, is that the ratio of microbial cells to human cells within an individual microbiome is around 10:1 (Luckey, 1972) but recent estimates disprove this myth and estimate the ratio to be around 1.3:1 (Sender et al., 2016a, 2016b) and highly variable between subjects; percentages as roughly 56.5% microbial and 43.5% human cells on average (Fig. 1). Although the ratio of microbial to human cells is closer to 1:1, there are an estimated 500-1000 species of bacteria that exist in the body at one time – though, the number of unique sub-species and genotypes may be magnitudes larger -

making the number of microbial genes vastly greater than the number of human genes (Gilbert et al. 2020, Turnbaugh et al., 2007, Locey et al. 2016); thus, dramatically increasing the metabolic heterogeneity and complexity of the ecosystem. Furthermore, the dynamics within a human microbiome is especially complex because of the influence from host genetics and environmental stimuli that introduce latent effects rendering disease-related dysbiosis a non-trivial topic to study. Characterizing the mechanisms of these interactions is paramount in the quest to understand health and dysbiotic systems.

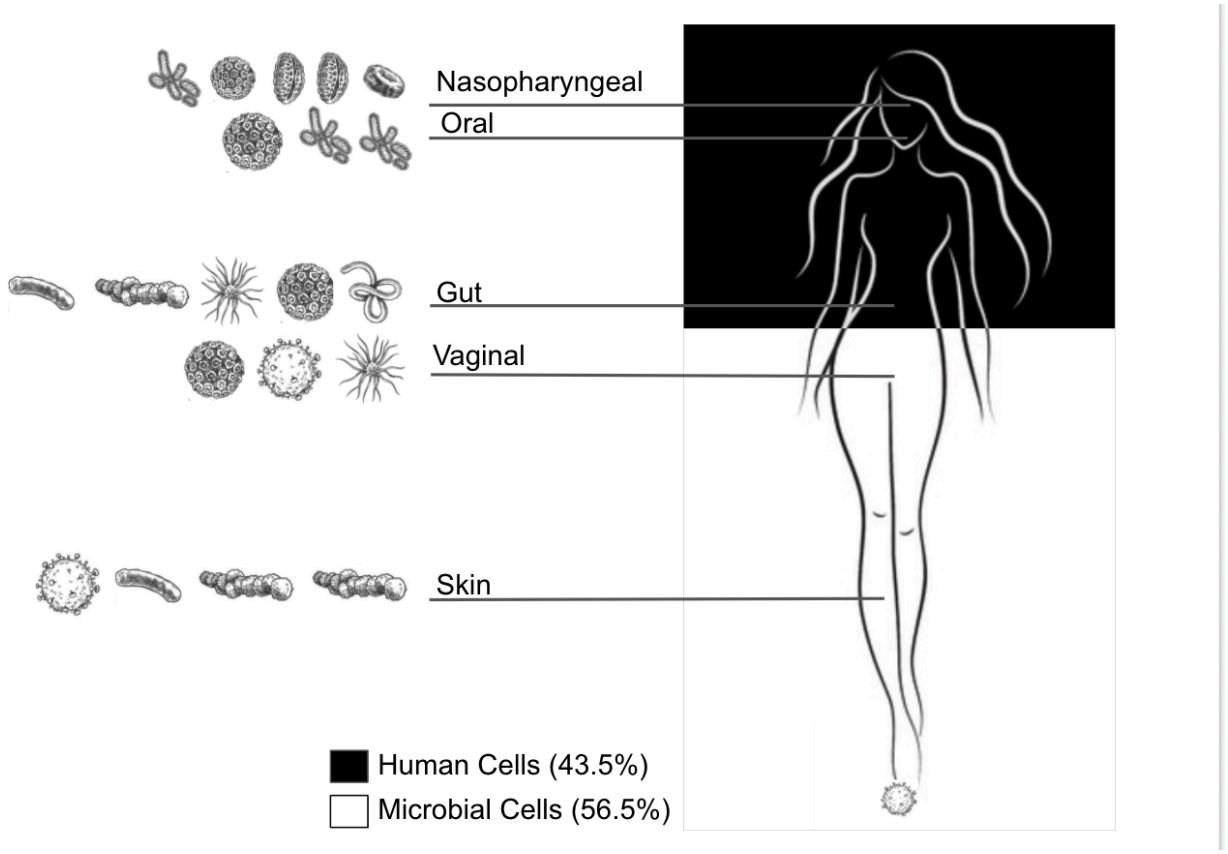


Figure 1 – Ecosystems of the human microbiome

(Right) Showcasing the 1.3:1 ratio of microbial to human cells proposed by (Sender et al., 2016a, 2016b) but represented as percentages (56.5% microbial, 43.5% human). (Left) Simplified illustration of varying communities and microbial richness for each ecosystem in the human microbiome consisting of, but not limited to, nasopharyngeal, oral, gut, vaginal, and skin.

2.2 – Next-generation sequencing technologies

With the advent of next-generation sequencing (NGS) technologies, deep profiling of biological systems has become increasingly affordable, and the collection of publicly available datasets is growing exponentially. NGS instruments estimate the relative abundance of discrete biological components (e.g., transcripts, 16/18S rRNA gene, marker genes) within a community by sampling from a pool of nucleic acid fragments. NGS technologies are the most practical avenue in which to study microbiomes and is harnessed via metagenomics and metatranscriptomics, collectively referred to as "omics" but the term encompasses other technologies, such as proteomics, as well. The cost of NGS technologies have fallen drastically since the first human genome has been sequenced with ~$10,000 per megabase of DNA sequence in 2001 to ~$0.01 in 2020; substantially lower than the predictions of ~$10 per megabase via Moore's law (NHGRI, n.d.). The US National Institute of Health (NIH) sequencing read archive alone contained more than 36 petabytes of raw NGS reads in 2020 and is expected to rise to 43 petabytes in 2023 (National Center for Biotechnology Information, 2020). From falling costs and rising accuracy of NGS technologies to the expanse of public data repositories, genomics and meta-omics have moved from the being used primarily by researchers to being increasingly used in clinical settings to investigate human and microbial-related diseases, respectively.

A major advantage of metagenomics and metatranscriptomics is the ability to sample from a variety of systems (e.g., marine systems, mammalian microenvironments, build

surfaces) in an untargeted way to study organisms that cannot (yet) be cultured in the laboratory. Early microbiology studies were cognizant to the prevalence of microbes that could not be cultured. For instance in 1932, Razumov reported discrepancies between cultivation-based and microscope-based cell counts (Razumov, 1932). Recently, it has become clear that most bacterial phyla lack cultured representatives (Hug et al., 2016); begging the question of what fraction of microscopic cells and viral particles on Earth have been cultured? While each environment has its own characteristics and are likely extremely variable, one study estimates that only about 0.01-1% of the microorganisms observed in a microscope could be isolated using artificial media (Garza and Dutilh, 2015). The ability to sequence and reconstruct genomes without the need for exhaustive culturing experiments not only expands our reference databases but more importantly widens our understanding of life while increasing the potential for identifying genes that can be repurposed via biotechnology.

The basic approach to NGS sequencing is isolation of genetic material (DNA or RNA) with subsequent sequencing of fragments to produce reads. The reads output by NGS instruments are then assembled *in silico* to produce contiguous sequences by taking these short nucleic acid sequences and combining them to reconstruct the original source genetic sequence. As NGS technologies become more widespread in their usage, assemblers have followed suite with the emergence of assembly algorithms that specialize in metagenomes (Nurk et al., 2017), transcriptomes (Bushmanova et al., 2019), biosynthetic clusters (Meleshko et al., 2019), plasmids (Antipov et al., 2016), viruses (Antipov et al., 2020), and even SARS-CoV-2 corona viruses (Meleshko et al., 2021);

many of which are based on a data structure called a de Bruijn graph (Compeau et al., 2011).

Two common types of metagenomic sequencing used in environmental and microbiome sampling experiments are marker gene surveys and shotgun metagenomic sequencing (Table 1). Both types can be referenced based but while the former requires a reference database, the latter can be post-processed as *de novo* to build a reference. Similar to shotgun metagenomics, there exists metatranscriptomics which can be processed as either reference based (often paired with *de novo* shotgun metagenomics) or *de novo* to yield transcripts that can be used as a reference. Although in both shotgun metagenomics and metatranscriptomics *de novo* approaches the output is often annotated with existing reference databases, they are still able to identify sample-specific genomes and metabolic activity not previously observed. Marker gene surveys, shotgun metagenomics, and metatranscriptomics all have their benefits and caveats which is why meticulous experimental design is critical to maximize the amount of resources devoted to researching a topic of interest.

The aim of marker gene surveys is to quantify the abundance of specific components and, by extension, the biological features they represent. Marker gene surveys amplify a specific region of the source genetic material, with the most common being the 16S ribosomal RNA (rRNA) for prokaryotic, 18S rRNA for protists, and ITS for fungal targets. The reads from this approach are aligned and clustered, often at 97% nucleotide identity, into operational taxonomic units (OTU), or amplicon sequence variants (ASV) where

unique sequences retain their independence. Although these surveys are cost effective with established user-friendly post-processing computational pipelines, they are limited to only characterizing taxonomy, often only at a genus specificity, and provide no functional information. Although the abundance of specific taxa may be the suitable for many studies, especially when the sample size is large, the lack of defined genomes in marker gene surveys makes it difficult (if not impossible) to distinguish between similar strains (Espinoza et al., 2018) and provide no means for characterizing novel organisms since analysis is dependent on existing databases. There are tools such PICRUSt2 that infer metabolic function from marker gene surveys but these predictive methods are only useful in well characterized systems such as the human gut and should not be used in less characterized systems such as environmental samples.

The aim of shotgun metagenomics is to not only quantify which taxa are present in a community as in marker gene surveys but to quantify the functional potential of these organisms. The library preparation differs from marker gene surveys as shotgun metagenomics protocols fragment the source genetic material and sequences all the DNA present in the sample instead of a specific locus. Shotgun metagenomics are often split into two categories: (1) reference-based profiling metagenomics; and (2) *de novo* assembly-centric. Profiling approaches are more user-friendly and typically implemented by mapping reads to a database of well-established marker genes to determine the taxonomy and, sometimes, metabolic potential (Beghini et al., 2021; Nayfach et al., 2016; Wood et al., 2019). *De novo* assembly-centric approaches are more complex as they assemble reads into contiguous sequences referred to as contigs and these contigs are

then joined into larger scaffolds. Although these methods are time consuming and analytically challenging, they provide much richer information; more than is necessary in some studies. As the main objective of assembly-centric shotgun metagenomics is typically to associate metabolic function with specific, post-processing analysis is non-trivial because the output data are random and have no biological organization. The post-processing of assembly-centric shotgun metagenomics, such as reconstructing genomes in a process called genome binning (described below), often requires more expertise in running computational pipelines and interpretation, which makes automation and scalability of shotgun metagenomics more difficult than marker gene surveys.

While metagenomics focuses primarily on identifying which microbes are present within a community and their functional potential, metatranscriptomics can be used to study the expression of genes within such communities and how these levels change in different conditions (e.g., healthy vs. dysbiotic microbiomes). Metatranscriptomics is more difficult than metagenomics approaches because it involves RNA which is less stable than DNA. A typical shotgun metatranscriptomics protocol includes RNA extraction, depletion of rRNA, and cDNA synthesis to prepare the libraries prior to sequencing. Similar to shotgun metagenomics, metatranscriptomics can either be reference-based profiling based or *de novo* assembly-centric. Metatranscriptomics can also be paired with *de novo* metagenomics, where the metatranscriptomic reads are mapped to genomes derived from metagenomic assemblies to determine gene expression. With regards to profiling approaches, many of the tools designed for microbial profiling can be used seamlessly with either RNA or DNA reads. In a *de novo* assembly-centric metatranscriptomic

approach, the reads are assembled into contiguous sequences as in shotgun metagenomic assembly but the resulting contigs represent transcripts that are organized by predictive *in silico* gene assignments. The advantage of metatranscriptomics is that it can provide information about differences in the functional activity of microbial communities which appear to be the same in terms of microbe composition (Bashiardes et al., 2016). In many "omics" studies, metagenomics and metatranscriptomics sequencing are paired where genomes are derived from the assembly-centric shotgun metagenomics and the reads from the metatranscriptomics are used to determine gene expression for these *de novo* communities as mentioned previously.

Table 1 – Overview of types of NGS approaches for quantifying environmental DNA and RNA

Type	Cost	Scalability	Reference-based	Taxonomy	Metabolism	Analysis Difficulty
MGX - Marker Gene Survey	*	***	✓	✓		*
Shotgun MGX/MTX (Profiling)	**	***	✓	✓	✓	*
Shotgun MGX (De novo)	**	*		✓	✓	***
Shotgun MTX (De novo)	**	**		✓	✓	**
Paired MGX - MTX (De novo)	***	*		✓	✓	***

MGX Metagenom cs
MTX Metatranscr ptom cs
MAG Metagenome Assemb ed Genome

2.3 - Reconstructing genomes from metagenomes

The methodology for reconstructing genomes *in silico* from metagenomic assemblies is referred to as genome binning. The ability to organize disjoint sequences into individual genomes is a major strength employed by shotgun metagenomics and is absolutely essential when linking functional potential to taxonomy; especially in the context of novel organisms. The minimum requirement input for running a genome binning algorithm is

the metagenomic assembly in the form of contigs or scaffolds. In most genome binning algorithms, the contigs are binned by making use of DNA sequence composition and counting the number of overlapping occurrences of *k*-mers, which are genetic sequences of length *k*. However, many algorithms take in additional input arguments including coverage matrices linking sequencing depth per contig per sample (Alneberg et al., 2014; Kang et al., 2019; Nissen et al., 2021; Wu et al., 2016), Hidden Markov Models identifying which marker proteins occur in each bin (Wu et al., 2016), codon usage (Yu et al., 2018), and even the de Bruijn graph intermediates of the metagenome assemblies (Mallawaarachchi et al., 2020, 2021). The input data are combined and clustered (a form of unsupervised machine learning described later) using various algorithms such as Emergent Self-Organizing Maps (Dick et al., 2009), t-Distributed Stochastic Neighbor Embedding (Van Der Maaten, 2014), Uniform Manifold Approximation and Projection (McInnes et al., 2018), or even deep learning techniques such as Variational Autoencoders (Nissen et al., 2021) to produce groups of contigs referred to as **genome bins**. Once the genome bins have been refined (Sieber et al., 2018), the quality has been evaluated (Parks et al., 2015), quality guidelines are met (Bowers et al., 2017)), and taxonomy has been determined (Chaumeil et al., 2020) the draft genomes are then referred to as metagenome assembled genomes (MAG). These genome binning techniques make it possible to recover genomes from organisms that cannot be cultivated and have broadened our range of detecting and characterizing microbial life from various environments (Hug et al., 2016). While this has been historically focused on Bacteria or Archaea, recent advancements in bioinformatics have made it routine to assemble viral genomes from metagenomes (Antipov et al., 2020), isolate complete or partial viruses

from assemblies (Guo et al., 2021; Kieft et al., 2020; Ren et al., 2017), and evaluate the quality of detected viruses (Nayfach et al., 2020).

2.4 – Next-generation sequencing data is inherently compositional

While NGS technologies prove invaluable in the quest for identifying and characterizing microbes that cannot yet be cultured, they are not without their caveats. One of the of the most obvious is the inability for short sequence reads to be resolved over repeat regions that are common in eukaryotic organisms. Emergent long read technologies such as Oxford Nanopore and PacBio can address these issues but are not without their flaws and not widely applied in meta-omics yet, thus they will not be further discussed in this dissertation.

Another caveat that is less known originates from the inherent design of NGS instruments as they estimate the relative abundance of discrete biological components (e.g., genomics fragments, transcripts, marker genes) within a community by sampling from a pool of nucleic acid fragments. Each of these subsampled fragments serve as a proxy for discrete biological components, and the percentage of each biological component is proportional to the true abundance in the sampled community. In practice, these subsampled fragments (i.e., reads) are mapped to a reference and the number of reads that align to a feature within the reference are counted resulting in a matrix (samples vs. features) of positive integers called a counts table. “Omics” counts tables, albeit routine in biological sciences, are non-trivial to analyze when considering the variance that arises from biological complexity, batch effects from laboratory protocols/conditions, and the

calibration of highly sensitive instruments. In addition to this inherent complexity, the data generated from NGS instruments is fundamentally compositional; that is, conveying only relative information dependent upon the capacity of the instrument, experimental design, and technical bias.

Compositional data is defined as a *D*-part composition, referred to as a composition, when all components are represented by strictly positive real numbers that carry only relative information (Aitchison, 1982; Egozcue et al., 2003). The statistical methods for compositional data address the reality that these compositions do not exist in real Euclidean space but in a subset known as the simplex (Aitchison, 1982). In NGS studies, these compositions are generated from subsampling and the observed abundance is uninformative (e.g. sequencing depth) as only this represents parts of a whole and carries only relative information (Pawlowsky-Glahn et al., 2015).

With this relativity, the information contained in the relationships between components is more essential than the information contained within an individual component (Rivera-Pinto et al., 2018). For instance, consider two conditions (A and B) where each composition contains 3 components (X, Y, and Z) where each component represents an individual microbe (Fig. 2, Table 2). Components Y and Z have identical abundance in both conditions but component X has doubled in condition B relative to condition A (i.e., $Condition_A(X)$ = 3, $Condition_B(X)$ = 6). If relative abundance measures are used and compared between conditions, then it may appear that components Y and Z are depleted in condition B relative to condition A but, in actuality, their absolute abundances have not

changed. Comparing relative abundances between samples will likely lead to spurious and noisy conclusions with faulty biological interpretations.

Log-ratio transformations are compositionally valid data transformations that perform equivalently on both the counts and relative abudances while capturing the relationships between features within the sample space (Pawlowsky-Glahn et al., 2015). The log-ratios are referred to as balances and are not sensitive to library-size or individual components. Aitchison geometry provides methods for analyzing compositional datasets through log-ratios and isomorphisms that transform compositions from the Aitchison simplex to real space (Aitchison, 1982). One type of Aitchison geometry is the isometric log ratio (ILR) which operates on relative data in an unconstrained space with an orthogonal basis, thus, preserving all information in the original composition and is preferred when a non-singular covariance matrix is needed (Egozcue et al., 2003; Silverman et al., 2017). The ILR method uses a sequential binary partition to construct a new set of coordinates making it desirable in the field of microbiology where phylogenetic trees represent a natural coordinate system for vertical evolutionary relationships that distinguish taxonomy-derived components. As mentioned, ILR transformations have recently been repurposed by the microbiology community and have proven to evade many statistical artifacts introduced from an incorrectly represented sample space (Morton et al., 2017; Silverman et al., 2017; Washburne et al., 2017). The ILR transformation has been used to investigate taxonomic signatures in the human gut that are associated with obesity (Finucane et al., 2014), has been benchmarked for several supervised machine learning methods against popular normalization techniques (Knights et al., 2011; Silverman et al.,

2017), and for linking external covariates to specific clades using regression methods (Washburne et al., 2017). ILR can also be used in a supervised setting for identifying taxa associated with a particular phenotype (i.e., feature selection) (Rivera-Pinto et al., 2018). However, one caveat of using balances is that the resulting data dimensionality is projected into a $D - 1$ dimensional space making interpretation difficult to directly associate specific components with particular trends.

Additive log-ratio transformation (ALR) also projects the data into a $D - 1$ dimensional space and does not preserve distances as it is an isomorphism but not an isometry (Pawlowsky-Glahn et al., 2011). However, ALR is not as common in biological settings as it often requires a single unchanged reference (Quinn et al., 2018), which is rarely available in many microbiome studies. Another type of Aitchison geometry is the centered log-ratio (CLR) transformation which preserves distances and is both an isomorphism and an isometry (Pawlowsky-Glahn et al., 2011). The CLR transformation is computed by taking the logarithm of each measurement and dividing by the geometric mean of the composition (i.e. arithmetic mean of logs) (Aitchison, 1982). An attractive feature of the CLR transformation is that the output retains the same dimensionality after transformation (Egozcue et al., 2003), which is not the case for ILR or ALR. This property allows for direct associations between a particular component and the transformed value without decomposing the balances amongst binary partitions as is required by ILR. CLR transformations have been applied to a wide range of biological topics including metagenomic binning using *k*-mer profiles (Laczny et al., 2015), the impact of gliadin in gluten-tolerant hosts (Zhang et al., 2017) and differential abundance (Fernandes et al.,

2013; Mandal et al., 2015; Morton et al., 2019). However, it is important to note that the CLR transformation yields a coordinate system featuring a singular covariance matrix which may violate the assumptions of some statistical methods (Pawlowsky-Glahn et al., 2015).

CLR- and ILR-transformed data benefit from the following properties: (1) scale invariance, in that multiplying by a constant, such as library-size, will not influence the resulting transformation; (2) perturbation invariance, in that converting compositions between equivalent units does not affect the results; (3) permutation invariance, in that the order of components comprising the composition does not matter; and (4) sub-compositional dominance detailing that a subset of a complete composition contains less information than the whole composition (Quinn et al., 2018).

Exploration in this landscape requires vigilance and awareness of data characteristics, such as the lack of independence among compositional features, when applying statistical methods not designed for such assumptions as described in the literature (Gloor et al., 2017; Quinn et al., 2018). Although many industry standard statistical approaches used for analyzing NGS datasets such as negative-binomial dispersion models (Love et al., 2014; Robinson et al., 2010) and zero-inflated gaussian models (Paulson et al., 2013) harbor assumptions about data distributions that violate the principles of compositionality (Gloor et al., 2017), the knowledge transfer of CoDA has migrated from geology to the realm of bioinformatics (Erb and Notredame, 2016; Lovell et al., 2015; Morton et al., 2017; Quinn et al., 2018). This facet of compositionality is

integral to CoDA, which has a strong mathematic theoretical foundation, and has become expected by many research groups that specialize in microbiome research. The aftermath of establishing CoDA principles as a cornerstone of NGS analysis has produced advancements in machine learning (Espinoza et al., 2021), phylogenetic analysis (Morton et al., 2017; Rivera-Pinto et al., 2018; Silverman et al., 2017; Washburne et al., 2017), network analysis (Erb and Notredame, 2016; Espinoza et al., 2020; Quinn et al., 2017), differential abundance analysis (Fernandes et al., 2014, 2013; Mandal et al., 2015; Morton et al., 2019; Thomas P. Quinn et al., 2018), metagenomic binning (Laczny et al., 2015), and feature engineering (Quinn and Erb, 2020).

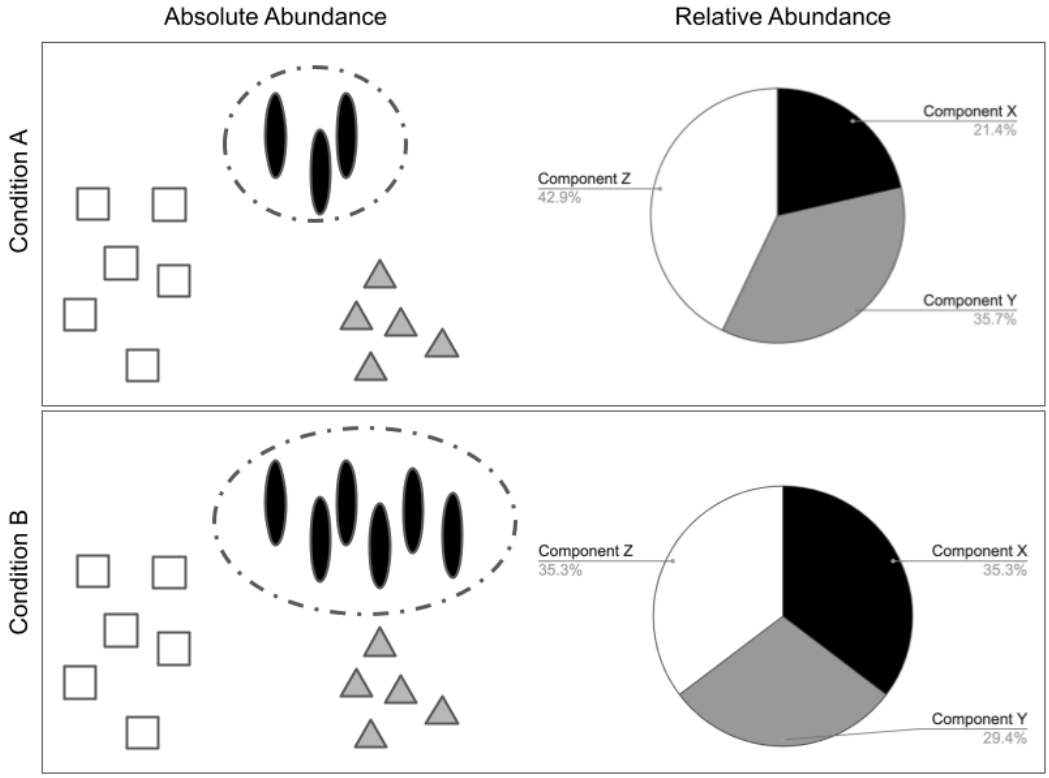


Figure 2 – Absolute vs. relative abundance of compositional data

Simplified illustration of absolute abundance (left) and the effects on relative abundance (right) for conditions A (top) and B (bottom) where component X is doubles from 3 to 6 counts.

Table 2 – Interpreting enrichment when comparing absolute and relative abundance between two conditions

Type	Condition	X	Y	Z
Absolute	A	3	5	6
	B	6	5	6
	∆ (B – A)	**3**	**0**	**0**
Relative	A	21.4%	35.7%	42.9%
	B	35.3%	29.4%	35.3%
	∆ (B – A)	**13.9%**	**-6.3%**	**-7.6%**

2.5 – *Structuring abstract associations in compositional data with networks*

A network is a graphical structure used to represent relations between discrete objects. It is a flexible abstract mathematical construct that can model systems with varying levels of complexity from simple binary networks to hierarchical weighted networks. This versatility is alluring to researchers seeking to understand how discrete features are associated with each other, reflective of the inner mechanics of a system. However, the versatility of networks theory comes with a cost in that a network is highly sensitive to input data, thresholds, inference, and transformations; therefore, implementation must be pursued strategically.

The discrete objects within a network are referred to as nodes and the connections between these nodes are referred to as edges (Fig. 3A). Edges are typically either weighted ($\boldsymbol{w} \in \mathbb{R}$ **)** or unweighted ($\boldsymbol{w} \in \{0,1\}$), where the weights correspond with some numeric association value that represents the connection strength between the two nodes; though, many layout algorithms expect networks with positive real valued edge weights (Jacomy et al., 2014). A matrix consisting of these values is called an adjacency matrix (***A***) and is the core structure of a network (Fig. 3B). ***A*** is often represented as a

(***m*** x ***m***) matrix where each $\boldsymbol{A_{ij}}$ represents a (weighted) connection defined by a real valued association function of two vectors of size ***n*** (***m*** = number of nodes and ***n*** = number of observations). Weighted networks are generally preferred over unweighted networks as they contain much more information than their binary counterpart.

Networks can either be symmetric or asymmetric, where the edges are undirected ($\boldsymbol{A_{ij}}$ equals $\boldsymbol{A_{ji}}$) or directed ($\boldsymbol{A_{ij}}$ may or may not equal $\boldsymbol{A_{ji}}$), respectively. Directed networks have been applied to modeling metabolic pathways, where nodes represent compounds and edges represent chemical reactions that transform metabolites into products (Levering et al., 2017). Undirected networks are more common when investigating compositional datasets and are covered in detail in the subsequent section of this review. As mentioned, the flexibility of networks is essential for its powerful applications while, often, interpretation is the limiting factor.

Network connectivity [k] is a metric used to quantify how much influence a particular edge, node, or group of nodes has within a system. Many studies benefit from weighted-degree as a connectivity measurement which is applicable at multiple levels and interpreted differently depending on the network context. Connectivity at the edge level represents the edge weight of a pair of nodes, while at the node level connectivity refers to the sum of weighted edges connected to a node (Fig. 3B,C). These connectivities can be grouped such as quantifying the total connectivity of a subset of edges (as in the case for node connectivity by grouping edges connected to a node), a subset of nodes, or the entire network itself. As scaled connectivity [$\tilde{k}$] normalizes the node connectivity values

so they sum to 1 within a network they can be used to compare networks with different numbers of nodes or edges (Fig. 3C).

Association networks are common amongst biological network analysis where each node typically represents a discrete feature, each edge represents an inferred interaction or association, and the edge weight represents the strength of association between a pair of nodes. Association networks construction is highly modular and customizable from the selection of the association metric to edge detection (Fig. 2). Common association measures include correlation coefficient (Fuller et al., 2007), -log(P) (Shomorony et al., 2020), mutual information (Lachmann et al., 2016; Villaverde et al., 2014), Kullback-Leibler divergence (Lachmann et al., 2016), and proportionality (Lovell et al., 2015). The most common perhaps is the correlation coefficient which measures the relationship strength between a pair of nodes and exists within the interval [-1,1] where a value of 1 indicates an identical relationship amongst covariates. There are several types of correlation measures including Pearson, Spearman, Kendall rank, and Biweight-midcorrelation. Pearson correlation measures linear relationships and is the most widely used correlation measure, albeit sensitive to outliers. Spearman correlation is a rank-based measure that is able to capture monotonic relationships while Biweight-midcorrelation is a median-based correlation. Both Spearman and Biweight-midcorrelation tests are more robust to outliers than Pearson, while the latter is often more powerful (Hardin et al., 2007; Langfelder and Horvath, 2012; Song et al., 2012). Correlation coefficients as an edge association metric are desirable because they are easily calculated, are subject to several asymptotic statistical tests, scaled, and the sign

of the measure can distinguish inverse relationships (Song et al., 2012). However, correlation measures can be biased by compositionality (Friedman and Alm, 2012).

Proportionality is a compositionally valid association measure that can be used as an analog to correlation as the range of values is from [-1,1]. There are typically three flavors of proportionality including ϕ, ϕ_s, and ρ_p (Lovell et al., 2015). The proportionality measures ϕ *and* ϕ_s both range between [0, ∞), as the asymmetric and symmetric versions, respectively (Quinn et al., 2017). The proportionality measure ρ_p is the most akin to correlation as the pairwise application results in a symmetric matrix with values ranging from [-1,1] where a value of 1 indicates perfect proportionality amongst components (Erb and Notredame, 2016). A major advantage of proportionality measures is that they are robust when analyzing relative data (Lovell et al., 2015) and tend not to produce spurious connections (Quinn et al., 2017); a stark contrast with Pearson's correlation coefficient which had considerable limitations when applied to compositional data. The properties of robustness to spurious results, scale invariance, and interpretability positions proportionality as an effective association metric when inferring cooccurrence (Bian et al., 2017) and coexpression (Lovell et al., 2015) from NGS-derived datasets.

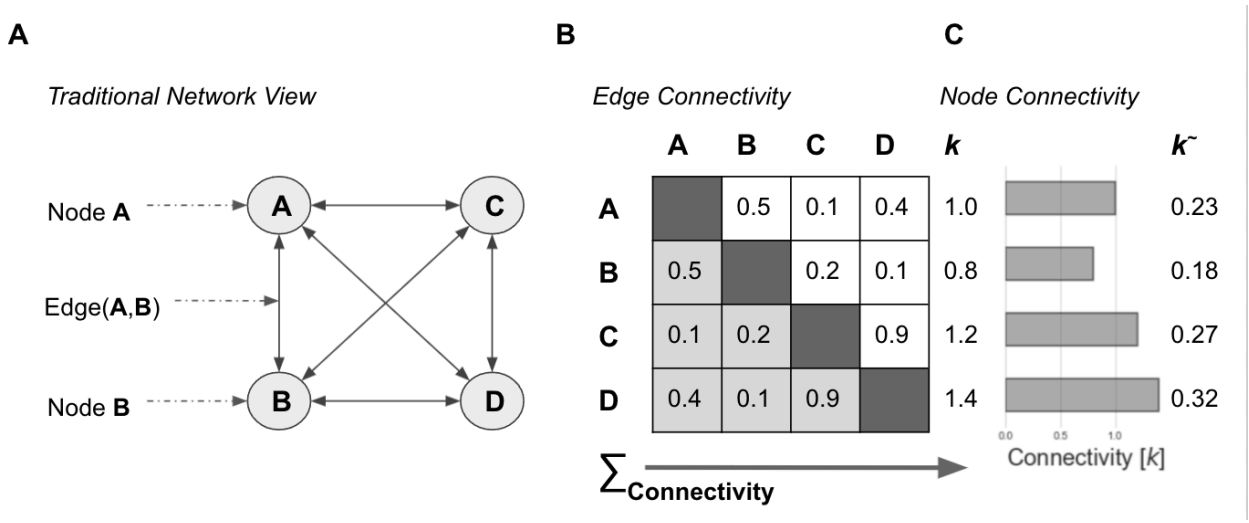


Figure 3 – Network connectivity

(A) Traditional view of a fully-connected network of 4 nodes {A,B,C,D} with undirected edges where Edge(A,B) = Edge(B,A). (B) Adjacency matrix representation of (A) edge connectivities showing symmetry of upper and lower triangles. In this scenario, self-associations are not included. (C) Node connectivity shown as weighted-degree, that is, the sum of edge weights connected to a node and scaled connectivity which normalizes the connectivity so that the total of all connections sums to 1.

2.6 – Machine learning as a subset of artificial intelligence

Artificial intelligence (AI) in popular culture and science fiction is much different in reality. In the most general sense, AI simply refers to programs with the ability to learn and reason as observed with human intelligence. However, this is partitioned into 2 categories: (1) universal AI; and (2) narrow AI. Often times in popular culture, the AI referenced is universal AI which exhibits all features of human intelligence but this type of AI research is much more niche and is not at the level of real world applications. Narrow AI on the other hand, though less well known amongst non-engineers, is much more applicable for

research purposes. Narrow AI can be defined as AI that exhibits some aspect of human intelligence and designed to accomplish a specific task very well.

Machine learning is a subset of narrow AI that provides systems with the ability to automatically learn and improve from experience without being explicitly programmed. The basic anatomy of input data in a machine learning algorithm includes the observations (e.g., biological samples), the features (e.g., genes), and, in the case of supervised learning, the target output (e.g., classifications or continuous values). Machine learning algorithms are typically split into 2 categories: 1) unsupervised learning; and 2) supervised learning (Fig. 4A). In NGS studies, a typical machine learning problem will consist of at minimum (1) a (***n*** x ***m***) feature matrix, denoted as ***X***, of compositional data for both supervised and unsupervised paradigms and (2) a ***n*** dimensional target vector, denoted as ***y***, in supervised paradigms though additional metadata (either with respect to observations or features) (Fig. 4B).

Unsupervised learning is a branch of machine learning in which the algorithm is not provided with any pre-assigned labels or scores for the training data (Baştanlar and Özuysal, 2014) and is typically used in clustering (e.g., genome binning) and network analysis. Supervised learning is another branch where algorithms learn a function that maps an input to an output based on example input-output pairs; that is, the algorithm infers a function from labeled training data. Supervised learning is split further into 2 subcategories with one being classification algorithms that predict sample labels and the other being regression algorithms that predict continuous values based on input data.

Consider a 2-dimensional dataset (dimensions *x* and *y*) with 2 classes of samples ($Class_A$ and $Class_B$) as illustrated in Fig. 4C where *x* and *y* are linearly related; this can either modeled as a regression or classification problem. For instance, if one were to model *y* as a function of *x* (i.e**.,** ***f(x) = y***), then the fitted model would be the line that best fits the trend of *x* with respect to ***y;*** the best-fit line as indicated by the dashed line in Fig. 4C. Similarly, this best-fit line can also be used a discriminator between $Class_A$ and $Class_B$ in a classification task where any point above the f*(x) = y* regression line can be classified as $Class_A$ and any point below the line as $Class_B$. However, most real-world datasets where the latent mechanisms giving rise to the phenomena being modeled are multifactorial and probabilistic (e.g., biology), the models are much more complex. Further, in many real-world datasets the number of samples within classes tend to differ as a result of poor experimental design, incomplete measurements (e.g., random sampling), or the (in)availability of *bona fide* samples within a class (e.g., compounds within a MOA), referred to as class imbalance illustrated in Fig. 4C, and this aspect tends to create biased models if not addressed properly.

In both scenarios, the true values of ***y*** would be compared against the model predictions, referred to as $\hat{\boldsymbol{y}}$ *,* to evaluate the performance of the model. In the regression scenario, a basic metric of performance would be mean squared error where the difference between $\hat{\boldsymbol{y}}$ and ***y*** are measured for each point, squared so all values are positive, and then averaged (i.e., MSE = Mean($(\hat{\boldsymbol{y}} - \mathbf{y})^2$)). In the classification scenario, the most basic metric would be accuracy which would get the number of correctly classified points divided by the number of total points (i.e., Accuracy = Mean($\hat{\boldsymbol{y}} == \boldsymbol{y})^2$).

The data that is used as input into the algorithms are typically split into training and testing datasets where the training set is used to fit the models and the testing set is used to evaluate the models (Fig. 4B). However, the training and testing is typically implemented in an iterative manner by randomly sampling training and testing pairs, given there isn't a hierarchical structure which is typically known *a priori* as described in **Publication I**, which prevents overfitting models and increases generalizability to new observations.

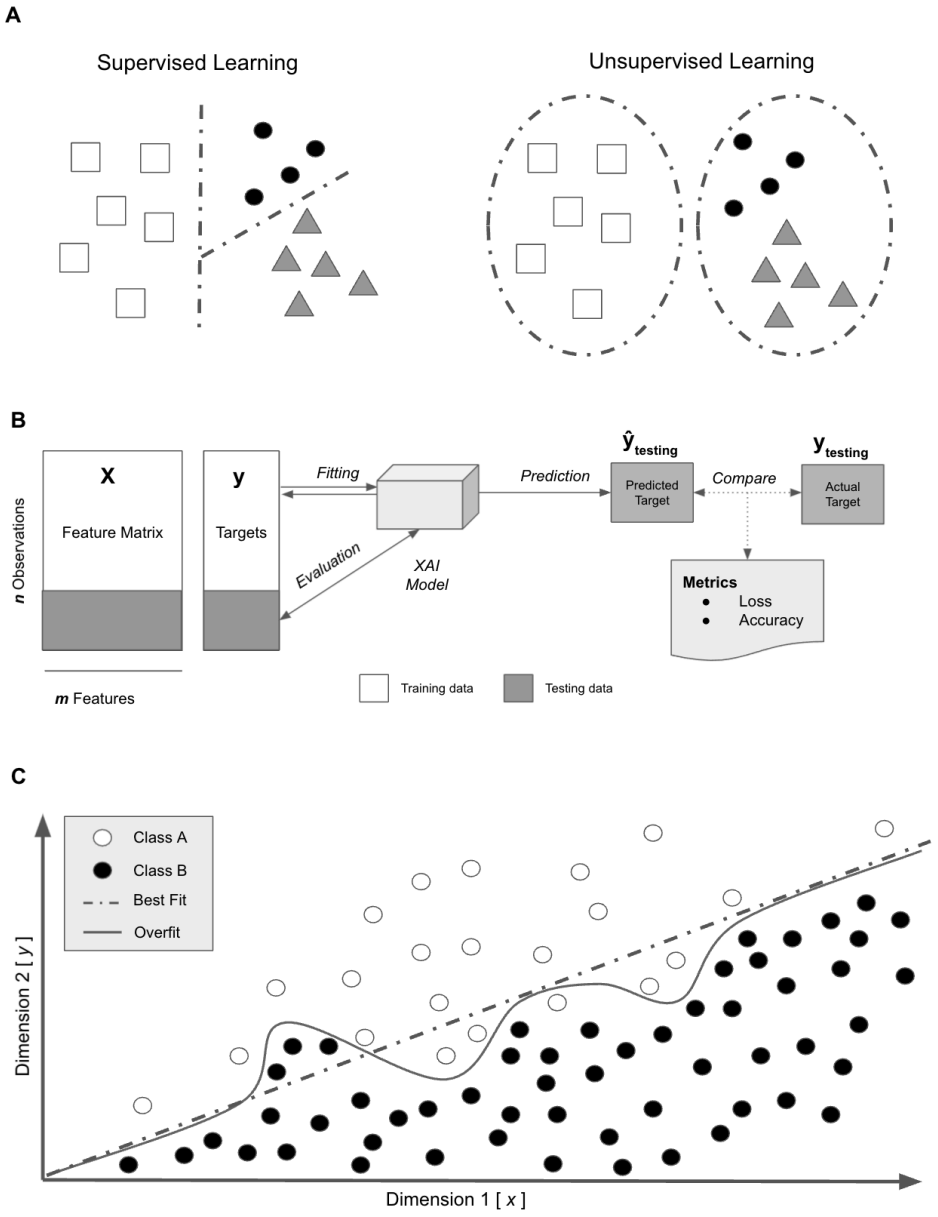


Figure 4 – Types of machine learning

(A) Visualizing the differences between supervised and unsupervised machine learning of samples representing different classes indicated by shapes. (A, Left) Supervised learning indicating best fit model classification sectors indicated by dashed lines separating samples of different classes. (A, Right) Unsupervised learning clustering samples learning patterns without any training *a priori* with cluster sectors indicates by dashed lines. (B) General process in which algorithms are trained and evaluated given input

data. (C) Visualization the difference between classification and regression learning while also showing how more complex models can overfit the data and remove the generalizability to new data.

2.7 – Explainable artificial intelligence

XAI is the intersection of domain expertise, mathematics, and computer science (Fig. 5A). XAI arises from the proper execution of the synergy between biology and machine-learning where a researcher: (1) creates human interpretable models without sacrificing performance (e.g., high prediction accuracy); (2) interprets the models to better understand the biology; (3) uses the enhanced interpretation biology to construct more realistic models.

The premise of XAI is that learned models can be unpacked and analyzed so that their contents have an interpretable structure. With this premise, there are several aspects of the learning process that must be taken into consideration: (1) the algorithm used for learning (e.g., logistic regression vs. neural network); (2) the data transformation used (e.g., center log-ratio vs. Box-Cox); (3) the evaluation metric and cross-validation procedure; and (4) feature representation procedure (e.g., aggregated features vs. principal components). When curating these aspects, the advantages of XAI is a feed-forward loop of knowledge where models can be interpreted by domain experts such biologists or chemists which can inform the engineer on how to develop more expert curated AI models (Fig. 5B).

The complexity of a machine learning model at times can often be associated with higher performance (Fig. 5C). However, greater complexity provides a higher probability of model overfitting if the model is designed poorly or the training data is not sufficient. For instance, revisit Fig. 4C from a classification perspective to notice the overfitted model (solid line) that has greater accuracy but at the cost of increased complexity and decreased the generalizability to data. Deep learning models such as neural networks can be very complex with activation functions and multiple internal layers of varying dimensionality rendering human interpretation difficult but not impossible. On the other end of the spectrum, there are rule-based algorithms that are not very generalizable. Linear models and decision trees are widely used as they have mediocre performance out-of-the-box but can easily implemented into ensemble approaches such as random forests, gradient boosting trees, or adaboost.

XAI is a universal effort, not limited to bioinformatics, to transition our methodology towards one of interpretability and high-performance (Gunning, 2017). The prospect of XAI in the quest to demystify the complexity of human microbial communities in the context of what it means to be healthy or diseased is immeasurable. Furthermore, the XAI methodologies developed to study human microbiomes can be used seamlessly to investigate sustainable agriculture (Singh et al., 2020), coral resilience to climate change (Santoro et al., 2021), build environments for public health (Weiss et al., 2018), space

travel (Voorhies et al., 2019), and global pandemics (Burchill et al., 2021).

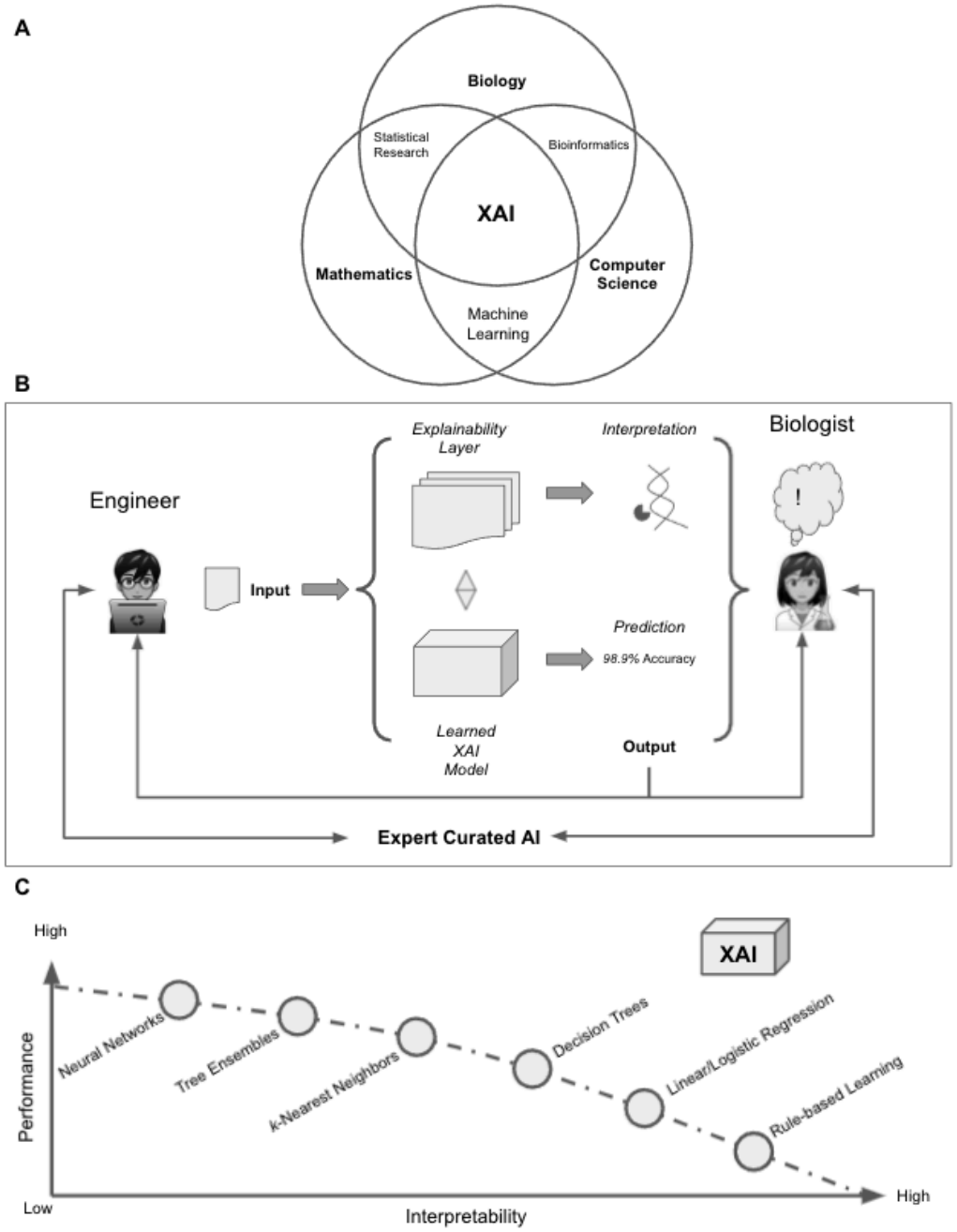


Figure 5 – Explainable artificial intelligence

(A) Venn diagram of the overlap between biology, mathematics, and computer science to yield XAI. (B) Schematic of the XAI learning process showing how models can be interpreted by domain experts in biology which in turn yield insight into how engineers can develop better models. (C) A diagram showing the general

relationship between a model's human interpretability and potential performance with XAI being in the sector that is both highly interpretable and highly performant.

2.8 - The union of domain expertise and artificial intelligence

Bioinformatics exists at an advantageous nexus between domain expertise and machine-learning with the requirement of framing biological questions from an engineering perspective and utilizing machine learning methods to address these questions.

Although NGS instruments output a deluge of data in the form of sequencing reads, the reference used for mapping these reads ultimately determines the resulting dimensionality of the counts table. In microbial isolate genomics, it is common to map the sequencing reads to ~4000 features (e.g., *Escherichia coli* K-12 genomes has 4,401 genes) and in human genomics there are typically ~20,000 features (more if non-coding genes or isoform variants are included), but in meta-omics it is not uncommon for a reference to contain over a million features. With most "Omics" datasets containing samples in the tens, sometimes in the hundreds and rarely in the thousands, this dimensionality has the potential to introduce a multitude of statistical anomalies if not properly addressed. More specifically, when the number of biological features vastly exceeds the number of biological samples statistical phenomena can arise that do not exist in lower dimensional representations (Altman and Krzywinski, 2018); a term dubbed as "the curse of dimensionality" (Bellman, 2003). The complexity, compositionality, and dimensionality of "Omics" datasets makes it challenging to derive clinical meaning and an understanding of the microbial system with respect to a host phenotype.

As the study of microbiomes become a more frequented vantage point in which to examine disease, new methods must be developed to surmount the obstacles inherent in the data. It is common for domain sciences such as biology to have analytics as the limiting factor for progress; the knowledge to ask the questions without the computational abilities to solve the problem on a large scale. This limiting factor is not only due to the complexity and dimensionality of the data but also the degree of uncertainty, unknown function of biological features, and influence from external variables. The quintessential example of this is illustrated in the synthesis of the minimal cell *JCV-syn3.0* with a 531 kilobase genome and 473 essential genes; 149 genes of which with completely unknown function (Hutchison et al., 2016). On the contrary, the fields of machine-learning and artificial intelligence have two limiting factors with the first being compute resources and the second being the domain expertise to interpret the findings in the context of domain knowledge; the means to solve the problem without knowing or understanding the question.

Naturally occurring microbial communities are structured by complex mechanisms including competition for resources, dynamic environmental conditions, and interactions between other organisms (Bastolla et al., 2009; Pascual-García et al., 2020). The study of systems biology is fundamental when investigating microbial-related diseases including what factors are involved in characterizing a healthy or diseased phenotype. Furthermore, the complexity of microbial communities introduces an unprecedented channel for discovering natural products as microbes wage a metabolic war when competing for limited resources. For instance, novel antibacterial compounds have been

isolated from soil microbiomes (Ling et al., 2015) and from the digestive systems of entomopathogenic nematodes (Imai et al., 2019). As these microbial communities cohabitate and influence their environment, antibiotic resistance becomes relevant in the diagnosis of microbial-related diseases as observed with drug resistance potential between the oral microbiomes of cohorts with and without dental caries (Espinoza et al., 2018)).

3 – PUBLICATIONS

Table 3 – Overview of peer-reviewed publications for dissertation

Publication	Type	Year	Journal	Impact Factor	DOI
I	Primary	2020	*Environmental Microbiology*	5.491	10.1111/1462-2920.15091
II	Primary	2021	*PLOS Computational Biology*	4.475	10.1371/journal.pcbi.1008857
III	Primary	2021	*EBioMedicine*	8.143	10.1016/j.ebiom.2021.103644
IV	Primary*	2022	*PNAS Nexus*	**	-
V	Secondary	2021	*Science Advances*	14.136	10.1126/sciadv.abg3088

Source: 2 year Impact factors determ ned v a *Academic Accelerator accessed 2021 11 18*

Primary type pub cat ons are d scussed n th s d ssertat on wh e *secondary* type pub cat ons are add t ona pub cat ons that use DUT aff at ons.

* Ind cates an add t ona pr mary pub cat on, not essent a for graduat on requ rements, that s n the peer rev ew process but cou d not be pub shed by the subm ss on of th s d ssertat on.

** *PNAS Nexus* s a new journa but s expected to have an mpact factor s m ar to PNAS

environmental microbiology

Environmental Microbiology (2020) **00**(00), 00–00 doi:10.1111/1462-2920.15091

Mini Review

Applications of weighted association networks applied to compositional data in biology

Josh L. Espinoza[1,2] **Naisha Shah**[1] **Suren Singh**[2] **Karen E. Nelson**[1,2,3] **and Chris L. Dupont**[1*]

[1]*J. Craig Venter Institute, La Jolla, USA.*

[2]*Applied Sciences, Durban University of Technology, Durban, South Africa.*

[3]*J. Craig Venter Institute, Rockville, USA.*

Summary

Next-generation sequencing technologies have generated, and continue to produce, an increasingly large corpus of biological data. The data generated are inherently compositional as they convey only relative information dependent upon the capacity of the instrument, experimental design and technical bias. There is considerable information to be gained through network analysis by studying the interactions between components within a system. Network theory methods using compositional data are powerful approaches for quantifying relationships between biological components and their relevance to phenotype, environmental conditions or other external variables. However, many of the statistical assumptions used for network analysis are not designed for compositional data and can bias downstream results. In this mini-review, we illustrate the utility of network theory in biological systems and investigate modern techniques while introducing researchers to frameworks for implementation. We overview (1) compositional data analysis, (2) data transformations and (3) network theory along with insight on a battery of network types including static-, temporal-, sample-specific- and differential-networks. The intention of this mini-review is not to provide a comprehensive overview of network methods, rather to introduce microbiology researchers to (semi)-unsupervised data-driven approaches for inferring latent structures that may give insight into biological phenomena or abstract mechanics of complex systems.

Received 23 March, 2020; revised 15 May, 2020; accepted 18 May, 2020. *For correspondence. E-mail cdupont@jcvi.org; Tel.: 8582001886; Fax: 8582001800

Introduction

With the advent of next-generation sequencing (NGS) technologies, deep profiling of biological systems has become increasingly affordable and the collection of publicly available datasets is growing exponentially. NGS instruments estimate the relative abundance of discrete biological components (e.g. transcripts, 16/18SS rRNA, marker genes) within a community by sampling from a pool of nucleic acid fragments. A typical NGS 16SrRNA experiment consists of (i) sequencing 16S rRNA amplicons at a specified sequencing depth from environmental samples (Caporaso *et al*., 2011; Logares *et al*., 2020), human tissues/biofluids (Goodrich *et al*., 2016; Gomez *et al*., 2017; Voorhies *et al*., 2019) or built environments (Weiss *et al*., 2018; Checinska Sielaff *et al*., 2019); (ii) clustering highly similar sequences into operational taxonomic units (OTU) as a representative for closely related taxa (Sneath and Sokal, 1962; Schloss *et al*., 2009; Edgar, 2013) or using amplicon sequencing variants (ASV) (Callahan *et al*., 2016; Amir *et al*., 2017); (iii) generating abundance tables by counting NGS reads mapped to ecological units (e.g. OTU or ASV); and (iv) analysing the resulting abundance tables such as alpha/beta-diversity, differential abundance or network analysis.

Each of these subsampled fragments serve as a proxy for discrete biological components, and the percentage of each biological component is proportional to the true abundance in the sampled community. However, the measured abundance ultimately depends on the chemistry of the assay, not the input material (Quinn *et al*., 2018), and observations of these biological units are not entirely independent as the instruments can only measure reads up to the capacity of the machine. For example, if we imagine a sequencer as having a fixed number of slots, analogous to sequencing depth, then an increased abundance of one biological component may saturate the available slots leaving fewer slots for less abundant components and

potentially rendering low abundance components as undetected. NGS experiments often produce millions of reads, and after quality control followed by post-processing, the resulting product is an abundance table organized by quantification of biological components with respect to the total measured reads within a sample represents the sequencing depth.

Although observed abundances in macroscopic ecological context are typically independent events (Gloor *et al*., 2017), this is often not the case in microbial ecology as many organisms cannot be cultured in a controlled setting (Rinke *et al*., 2013) and can currently only be observed collectively through NGS technologies. As a result of this phenomena, NGS-derived datasets such as metagenomics and metatranscriptomics are inherently compositional. Compositional data are defined as a vector of strictly positive real numbers with an unknown or uninformative total (e.g. sequencing depth) as the abundance of each component represents parts of a whole and carries only relative information (Pawlowsky-Glahn *et al*., 2015). With this relativity, the information contained in the relationships between components is more essential than the information contained within an individual component (Rivera-Pinto *et al*., 2018). Exploration in this landscape requires vigilance and awareness of data characteristics, such as the lack of independence among compositional features, when applying statistical methods not designed for such assumptions as described in the literature (Gloor *et al*., 2017; Quinn *et al*., 2018).

The advantages of powerful analytical techniques such as machine-learning and network analysis on NGS datasets become increasingly attainable as sequencing costs continue dropping by several orders of magnitude. For instance, in 2001, the cost of sequencing the human genome was ~$100,000,000 compared to ~$1000 in 2019; exceeding far beyond the Moore's law predictions of ~$180,000 (DNA Sequencing Costs, 2020), machine-learning methods are routinely applied on two-dimensional data matrices represented by observations and features, where each feature represents an individual measurable property or characteristic of a phenomenon being observed (e.g. microbiome diversity) (Bishop, 2006). This terminology adapts to compositional datasets where an observation would represent a particular composition, a feature as an individual component and the numeric value as the count of the component within the composition. Extending the concepts into network theory, these discrete features can represent nodes within a network and continuous associations between features as weighted edges. In the context of a NGS-derived compositional dataset, a typical network instance could entail an individual sample as an observation/composition, OTU as a features/component and read counts where pairwise association between features would be the basis for the edge connectivity within the network. Later in this review, we elaborate on the methods behind generating such networks from compositional datasets.

With large datasets, researchers are not only investigating the abundance/depletion of features in relation to a specific condition, but also the (inferred) interactions between features. One way for such an investigation is by applying network theory. The versatility of graphical abstractions using nodes, edges and topological structure can be contextually applied to a wide array of problems. For instance, applications of network theory have been successful in several fields including studying plankton networks driving carbon export (Guidi *et al*., 2016), gene interactions related to weight physiology (Fuller *et al*., 2007), ecological shifts (Gomez *et al*., 2017) and metabolic potential (Espinoza *et al*., 2018) associated with carious lesions in children and regulatory metabolic interactions in marine diatoms (Levering *et al*., 2017) and bacterial soil communities (Mandakovic *et al*., 2018). Many biological networks are composed of molecules such as DNA, RNA, proteins and metabolites as the nodes, and edges between these nodes represent either curated or inferred interactions between them. Furthermore, advanced multi-omics approaches incorporating associations across modalities such as clinical tests, proteomics, amplicon, transcriptomics, cytokines, metabolomics and lipidomics have begun to pave the way towards precision health using systems biology (Schüssler-Fiorenza Rose *et al*., 2019; Zhou *et al*., 2019; Shomorony *et al*., 2020). There are several approaches for network analysis in systems biology that each have their advantages and caveats. The goal of this review is not to describe the landscape of network methods but to guide the reader through the process of implementing association networks from NGS-derived datasets which are inherently compositional.

Compositional data

Compositional data is defined as a *D*-part composition when all components are strictly positive real numbers that carry only relative information (Aitchison, 1982; Egozcue *et al*., 2003). The statistical methods for compositional data address the reality that these compositions do not exist in real Euclidean space but in a subset known as the simplex (Aitchison, 1982). Datasets generated from NGS technologies such as gene expression and 16/18S amplicons are compositional, sparse and have complex distributions such as negative binomial (Robinson *et al*., 2010; Love *et al*., 2014), zero-inflated gaussian (Paulson *et al*., 2013) and Dirichlet (Holmes *et al*., 2012; Chen and Li, 2013; Wadsworth *et al*., 2017;

Harrison *et al*., 2019). Generating these datasets are often patient or sample limited making it difficult to produce a large number of sample observations without extensive resources and cooperation among collection agencies. This dilemma often subjects the dataset to 'the curse-of-dimensionality' in which the number of features vastly exceeds the number of observations potentially introducing statistical artefacts that can bias downstream analysis such as false positive correlations (Bellman, 2003). NGS-derived datasets are inherently incomplete as they are only parts of a complete system due to both biological and technical phenomena such as the capacity of a sequencing instrument to process reads (Gloor *et al*., 2017). This aspect presents a significant hurdle in analysis because a zero value may have different meanings in different datasets, and it is often difficult to distinguish the difference between true and false zeros (Kuhnert *et al*., 2005; Martin *et al*., 2005). Zeros can arise from many sources: (i) false zeros result from errors in experimental design or observational instrumentation; (ii) true zeros are either structural zeros and hypothesized in the statistical model or random zeros resulting from sampling variability (Parada *et al*., 2016; Blasco-Moreno *et al*., 2019). If not properly accounted for analytically, this missing information can introduce substantial artefacts in the downstream statistical analysis including comparing intra-sample patterns or association with other variables such as a phenotypic measurement. It is often useful to minimize excess sparsity to focus on core components within a system. The most common method for dealing with sparsity involves either filtering by prevalence, the addition of a minimal pseudocount or both respectively. However, incorporation of pseudocounts should be pursued with caution as haphazard usage can introduce statistical bias (Kumar *et al*., 2018). The handling of sparse data is an active area of research with interesting recent developments such as Robust Aitchison PCA (Martino *et al*., 2019). The balance between removing low prevalence features and retaining discriminative diversity depends on the research questions. For instance, if one is investigating community richness of soils in the context of potential natural products (Ling *et al*., 2015; Crits-Christoph *et al*., 2018), then it is reasonable to not remove any features assuming that the appropriate quality control and preprocessing were performed. In the scenario of inferred interactions in relation to a phenotype, it may be more beneficial to remove features with missing information to detail the relationships among core components. Most research questions will require a balance between the two extremes with some indication of when too many discriminative features are removed or when too many missing values have biased the data such as the notorious horseshoe effect (Diaconis *et al*., 2008; Morton *et al*., 2017).

The advantages and caveats of compositional data transformation

Relative abundance sensitivity and biased outcomes

Normalizations and transformations are standard approaches that are applied to compositional datasets when pursuing any type of weighted analysis beyond binary presence/absence and log-ratios. The most common normalization technique is total sum scaling (TSS), also referred to as relative abundance or closure (Aitchison, 1982), which divides each feature count by the sum of total counts in a sample. This technique removes technical bias that is related to differences in sequencing depth across samples. Despite the widespread adoption of this simple normalization, the abundances of specific components can drastically bias the results.

With TSS normalized data, the distance between variables is sensitive to the presence or absence of individual components and can reveal spurious relationships amongst unrelated variables resulting in false positive correlations (Pearson, 1896; Aitchison *et al*., 2000; Quinn *et al*., 2018). Consider the example illustrated in Fig. 1 involving a synthetic community of three OTUs with the following sample states: $sample_A$ (uniform abundances); $sample_B$ (doubling the abundances of OTU_1) and $sample_C$ (halving the abundances OTU_1). Notice the observed abundances of the community (Fig. 1A) and the TSS normalized abundances (Fig. 1B) show conflicting results when comparing between the samples. An increase in the abundance of OTU_1 within $sample_B$ introduces a false sense of depletion of OTU_2 and OTU_3, and the decrease in abundance in $sample_C$ suggests an enrichment of OTU_2 and OTU_3 when in reality their abundances did not change between samples. This artificial enrichment or depletion can lead to false positives in downstream analysis when investigating relationships between samples (e.g. network analysis, differential abundance, etc.). Using balance-trees is one method that bypasses the bias induced from TSS normalization as the ratios only reflect the values of the descendent nodes. A simplified example of balance trees is shown in Fig. 1C and D, where each internal node computes a ratio of the summed counts across both bifurcated paths. OTU_1 is a descendent of *y1* and not *y2*, and we illustrate that internal node *y1* is different between $sample_B$ and $sample_C$, whereas internal node *y2* is unchanged (Fig. 1C and D). This simplified concept of balance trees is vastly expanded in Aitchison geometry, in particular with isometric log ratio, and was first explored in geology (Egozcue and Pawlowsky-Glahn, 2005; Pawlowsky-Glahn and Egozcue, 2011). Since then, Aitchison geometry has recently been adopted by microbial-based data-science and is an effective alternative to address these

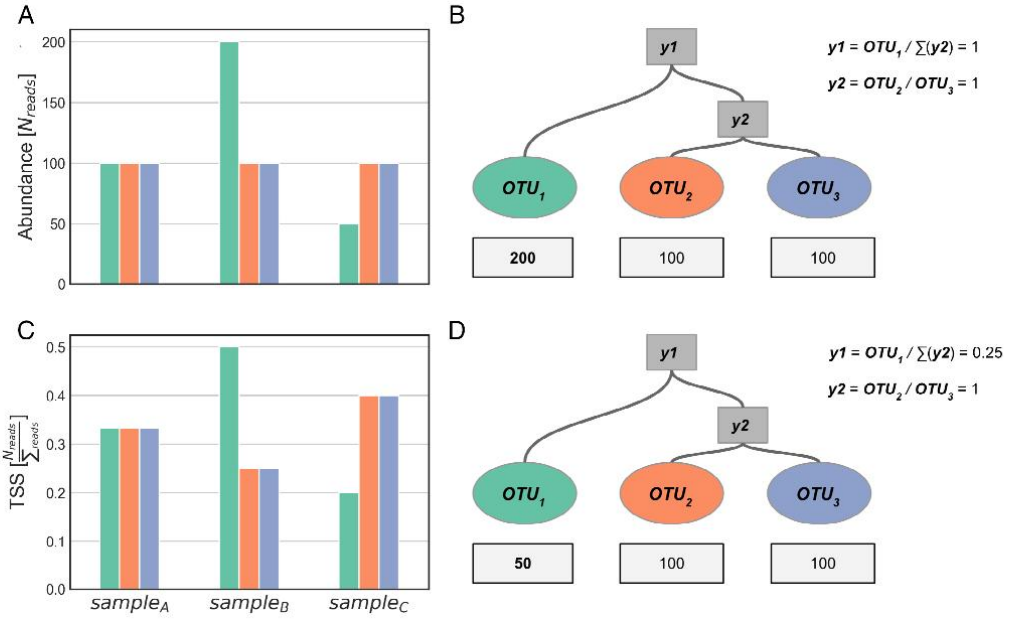


Fig 1. Comparison of balances and relative abundance with compositional data.A, B. Simple synthetic microbiome with three OTUs (OTU_{1-3}) and three samples ($sample_{A-C}$) represented as (A) absolute abundances and (B) relative abundances. C, D. Simplified examples of balance-trees where each internal node computes a ratio of the summed counts across both bifurcated paths for $sample_B$ and $sample_C$ respectively.

caveats (Silverman *et al*., 2017; Washburne *et al*., 2017; Rivera-Pinto *et al*., 2018; Morton *et al*., 2019).

Log-ratio transformations are a natural representation of compositional data

Log-ratio transformations perform equivalently on both the counts and proportions while capturing the relationships between features within the sample space (Pawlowsky-Glahn *et al*., 2015). The log-ratios are referred to as balances and are not sensitive to library-size or individual components. Aitchison geometry provides methods for analysing compositional datasets through log-ratios and isomorphisms that transform compositions from the Aitchison simplex to real space (Aitchison, 1982). One type of Aitchison geometry is the isometric log ratio (ILR) which operates on relative data in an unconstrained space with an orthogonal basis, thus, preserving all information in the original composition and is preferred when a non-singular covariance matrix is needed (Egozcue *et al*., 2003; Silverman *et al*., 2017). The ILR method uses a sequential binary partition to construct a new set of coordinates making it desirable in the field of microbiology where phylogenetic trees represent a natural coordinate system for vertical evolutionary relationships that distinguish taxonomy-derived components. As mentioned, ILR transformations have recently been repurposed by the microbiology community and have proven to evade many statistical artefacts introduced from an incorrectly represented sample space (Silverman *et al*., 2017; Washburne *et al*., 2017; Morton *et al*., 2017). The ILR transformation has been used to investigate taxonomic signatures in the human gut that are associated with obesity (Finucane *et al*., 2014), has been benchmarked for several supervised machine learning methods against popular normalization techniques (Knights *et al*., 2011; Silverman *et al*., 2017) and for linking external covariates to specific clades using regression methods (Washburne *et al*., 2017). ILR can also be used in a supervised setting for identifying taxa associated with a particular phenotype (i.e. feature selection) (Rivera-Pinto *et al*., 2018). However, one caveat of using balances is that the resulting data dimensionality is projected into a $D - 1$ dimensional space making interpretation difficult to directly associate specific components with particular trends. Additive log-ratio transformation (ALR) also projects the data into a $D - 1$ dimensional space and does not preserve distances as it is an isomorphism but not an isometry (Pawlowsky-Glahn *et al*., 2011). However, ALR is not as common in biological settings as it often requires a single unchanged referenced

(Quinn *et al*., 2018), which is rarely available, and will not be described further in this review. Another type of Aitchison geometry is the centered log-ratio (CLR) transformation which preserves distances and is both an isomorphism and an isometry (Pawlowsky-Glahn *et al*., 2011). The CLR transformation is computed by taking the logarithm of each measurement and dividing by the geometric mean of the composition (i.e. arithmetic mean of logs) (Aitchison, 1982). An attractive feature of the CLR transformation is that the output retains the same dimensionality after transformation (Egozcue *et al*., 2003), which is not the case for ILR or ALR. This property allows for direct associations between a particular component and the transformed value without decomposing the balances amongst binary partitions as is required by ILR. CLR transformations have been applied to a wide range of biological topics including metagenomic binning using *k*-mer profiles (Laczny *et al*., 2015), the impact of gliadin in gluten-tolerant hosts (Zhang *et al*., 2017) and differential abundance (Fernandes *et al*., 2013; Mandal *et al*., 2015; Morton *et al*., 2019). However, it is important to note that the CLR transformation yields a coordinate system featuring a singular covariance matrix which may violate the assumptions of some statistical methods (Pawlowsky-Glahn *et al*., 2015).

CLR- and ILR-transformed data benefit from the following properties: (1) scale invariance, in that multiplying by a constant, such as library-size, will not influence the resulting transformation; (2) perturbation invariance, in that converting compositions between equivalent units does not affect the results; (3) permutation invariance, in that the order of components comprising the composition does not matter; and (4) sub-compositional dominance detailing that a subset of a complete composition contains less information than the whole composition (Quinn *et al*., 2018).

Recently, the CLR transformation has been modified by multi-additive log-ratio (MALR) transformations to use the geometric mean of a subset of components as a specialized reference (Quinn *et al*., 2018). *Fernandes et al*. introduced the interquartile log-ratio (IQLR) transformation which includes only components that exist within the interquartile range of total variance in the geometric mean calculation implemented in their *ALDEx2* package (Fernandes *et al*., 2013, 2014). Another variant is the robust centered log-ratio (RCLR) transformation which uses only the non-zero components introduced by *Martino et al*. in their *DEICODE* package (Martino *et al*., 2019). Every log-ratio transformation uses a unique reference and applications of such will be dependent on the hypothesis. For instance, if an analysis does not require analysis of individual components such as beta diversity analysis (e.g. clustering, ordination, manifold learning) then ILR may be the most effective. If the desired outcome is to analyse individual components, then perhaps CLR or a MALR variant may the preferred option over ILR which would require traversing the basis partitions. The selection of which method CLR-based method will again depend on the preferred interpretation. If the query dataset is relatively tame in terms of outliers, then CLR would be a practical option as the reference would be based on the sample's geometric mean. If outliers were an issue, then it may be more reasonable to use IQLR where the reference is based on the geometric mean of the components in the interquartile range. Understanding the concepts behind the reference sets will allow researchers to continue extending these methods to more specialized applications and expanding the paradigm of compositional data analysis.

Principal component-based correction

Compositional datasets in biology can often contain anomalies derived from latent technical or biological phenomena. When uncorrected, these confounding effects can lead to spurious associations in network-based analyses described in detail below (Parsana *et al*., 2019). Latent factor-based data correction has successfully been applied to address variation introduced from batch effects (Goh *et al*., 2017) and can be extended to other sources of variance. Recently, *Parsana et al*. have developed methods to regress out latent confounders captured within the top principal components (Parsana *et al*., 2019). Principal component-based correction can be described by the following: (1) compute $\boldsymbol{p}$ principal component loadings ($\boldsymbol{L}$) where $\boldsymbol{p}$ reflects the number of principal components used during singular value decomposition; (2) fitting a linear model on each feature $\boldsymbol{E_i = \mu_i + \beta_i \times L_{1:p}}$ and (3) return the residuals. The assumptions of these methods posit that the network structure is scale-free (described below) and that the true topology is relatively sparse. The principal component-based correction implemented in *Parsana et al*. has been shown to reduce false positive associations when implemented as a preprocessing step before network construction using widely practiced methods such as weighted gene coexpression network analysis (WGCNA) (Langfelder and Horvath, 2008) and graphical lasso (Friedman *et al*., 2008) compared to instances with uncorrected input data. However, biological signals of interest may be removed by regressing out the top principal components in networks that violate the scale-free assumption. This technique could be adapted to incorporate non-Euclidean distances with alternative ordination methods and/or more complex manifold learning algorithms such *t*-Distributed Stochastic Neighbourhood Embeddings (t-SNE) (van der Maaten, 2013; Van Der Maaten, 2014) or Uniform Manifold Approximation and Projection (UMAP) (McInnes *et al*., 2018). Principal component-based correction, along with the proposed

adaptations, have substantial potential in correcting for biological and technical variance. Although these methods have been in practice for at least a decade (Price *et al*., 2006), they have not been extensively benchmarked and explored on the wide array of modalities produced by NGS technologies and present a unique avenue in which to interrogate noisy datasets. Aside from principal component-based correction, noise reduction can be implemented in the experimental design with custom mixtures consisting of mock microbial communities (Bokulich *et al*., 2016; Parada *et al*., 2016) or spike-in internal controls such as External RNA Controls Consortium (ERCC) standards composed of synthetic RNA oligonucleotides spanning a range of nucleotide lengths and concentrations (Pine *et al*., 2016). However, these later methods must be embedded in experimental design and cannot be utilized *post hoc*.

Network theory, metrics and applications

Overview of network structure and terminology

A network is a graphical structure used to represent relations between discrete objects. It is a flexible abstract mathematical construct that can model systems with varying levels of complexity from simple binary networks to hierarchical networks. This versatility is alluring to researchers seeking to understand how discrete features are associated with each other, reflective of the inner mechanics of a system. However, the versatility of networks theory comes with a cost in that a network is highly sensitive to input data, thresholds, inference and transformations; therefore, implementation must be pursued strategically.

The discrete objects within a network (**G**) are referred to as nodes or vertices (**V**), and the connections between these nodes are referred to as edges (**E**); formally, as ***G (V, E)***. Edges are typically either weighted ($\boldsymbol{w} \in \mathbb{R}$) or unweighted ($\boldsymbol{w} \in \{0,1\}$), where the weights correspond with some numeric association value that represents the connection strength between the two nodes; although, many layout algorithms expect networks with positive real valued edge weights (Jacomy *et al*., 2014). This is formulated as ***(i, j, w)*** where ***i*** and ***j*** represent the source and target nodes respectively and ***w*** indicates edge weight. This representation of a static network can be extended to a dynamic network such as a temporal network, where edges are described by ***(i, j, w, t)*** as ***t*** represents time. A matrix consisting of these values is called an adjacency matrix (***A***) and is the core component of a network. ***A*** is often represented as a ($\boldsymbol{m} \times \boldsymbol{m}$) symmetric matrix where each $\boldsymbol{A}_{ij}$ represents a (weighted) connection defined by a real valued association function of two vectors of size ***n*** (***m*** = number of nodes and ***n*** = number of observations). Weighted networks are generally preferred over unweighted networks as they contain much more information than their binary counterpart. The weighted networks can either be symmetric or asymmetric, where the edges are undirected ($\boldsymbol{A}_{ij}$ equals $\boldsymbol{A}_{ji}$) or directed ($\boldsymbol{A}_{ij}$ may or may not equal $\boldsymbol{A}_{ji}$), respectively. Directed networks have been applied to modelling metabolic pathways, where nodes represent compounds and edges represent chemical reactions that transform metabolites into products (Levering *et al*., 2017). Undirected networks are more common when investigating compositional datasets and are covered in detail in the subsequent section of this review. Bayesian networks, a type of directed network, have been used to investigate longitudinal relationships within the infant gut microbiome (Mcgeachie *et al*., 2016) but take much longer to compute than undirected association networks and will not be further covered in this review. As mentioned, the flexibility of networks is essential for its powerful applications, whereas, often, interpretation is the limiting factor.

Association measures

Association networks are common among biological network analysis where each node typically represents a discrete feature, each edge represents an inferred interaction or association and the edge weight represents the strength of association between a pair of nodes. Association networks construction is highly modular and customizable from the selection of the association metric to edge detection (Fig. 2). Common association measures include correlation coefficient (Fuller *et al*., 2007), –log(*P*) (Shomorony *et al*., 2020), mutual information (Villaverde *et al*., 2014; Lachmann *et al*., 2016), Kullback–Leibler divergence (Lachmann *et al*., 2016) and proportionality (Lovell *et al*., 2015). The most common perhaps is the correlation coefficient which measures the relationship strength between a pair of nodes and exists within the interval [−1,1] where a value of 1 indicates an identical relationship among covariates. There are several types of correlation measures including Pearson, Spearman, Kendall rank and Biweight-midcorrelation. Pearson correlation measures linear relationships and is the most widely used correlation measure, albeit sensitive to outliers. Spearman correlation is a rank-based measure that is able to capture monotonic relationships, whereas Biweight-midcorrelation is a median-based correlation. Both Spearman and Biweight-midcorrelation tests are more robust to outliers than Pearson, whereas the latter is often more powerful (Hardin *et al*., 2007; Langfelder and Horvath, 2012; Song *et al*., 2012). Correlation coefficients as an edge association metric are desirable because they are easily calculated, are subject to several asymptotic statistical tests, scaled, and the sign of the measure can distinguish inverse

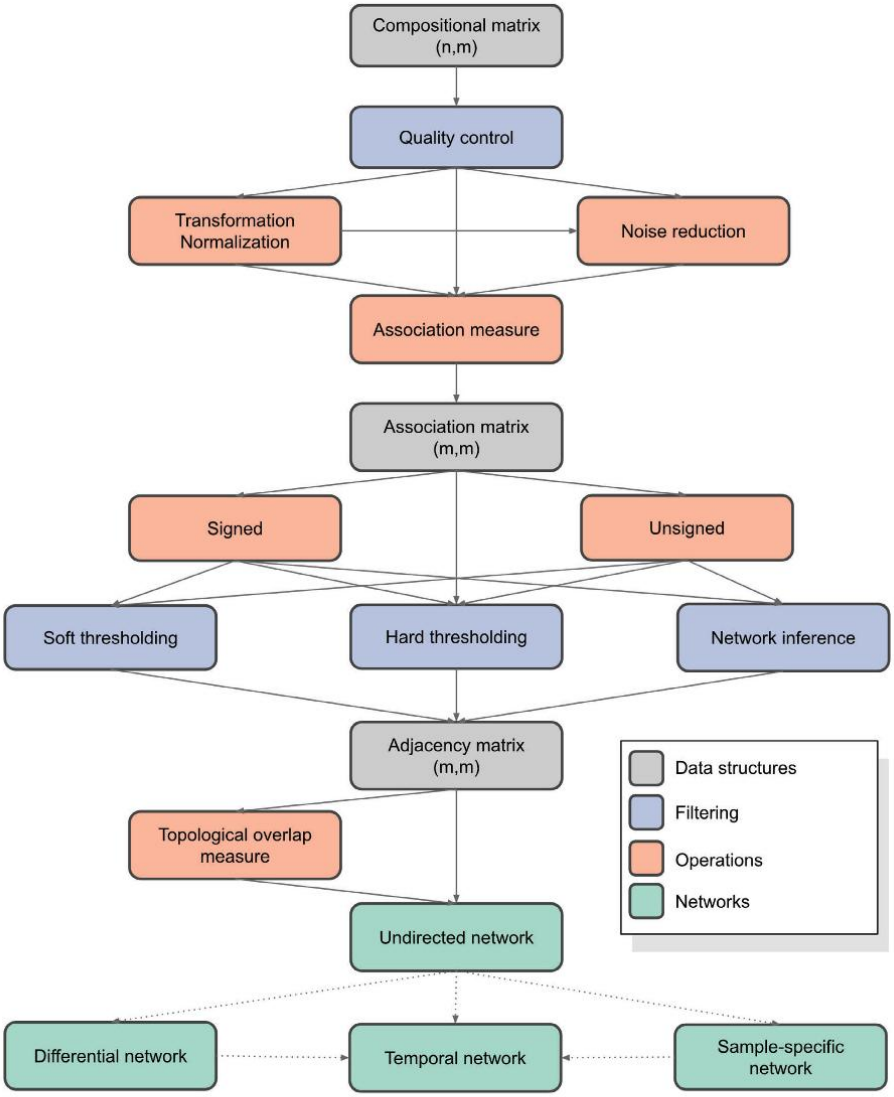


Fig 2. Schematic illustrating the modularity of network analysis.Modularity of network construction visualized by a network. The top of the visualization starts with compositional data as output from NGS methods. Traversing vertically goes further into the various operations detailed in the review to construct complex networks.

relationships (Song *et al*., 2012). However, correlation measures can be biased by compositionality (Friedman and Alm, 2012).

Many inference methods have been developed to leverage compositionality with sparse solutions to mitigate the effects of this bias including REBECCA (Ban *et al*., 2015), SparCC (Friedman and Alm, 2012), SPIEC-EASI (Kurtz *et al*., 2015) and CCLasso (Fang *et al*., 2015), whereas others rely on probabilistic graphical models (Tackmann *et al*., 2019) or permutation-based methods (Faust and

Raes, 2016). Variance log-ratio (VLR) is another compositionally valid association metric that does not produce spurious results. VLR measures the concordance between two compositions (e.g. samples, observations or feature vectors) and computes the variance of the logarithm of one component as divided by a second component (Quinn *et al*., 2018). However, VLR has a substantial limitation in that it is unscaled with respect to the variances of the log components with the range [0, ∞), where zero indicates perfect coordination (Aitchison, 1982).

Proportionality is another compositionally valid association measure, implemented in the *propr* package, introduced by *Lovel et al*. and expanded upon by *Erb et al*. (Lovell *et al*., 2015; Erb and Notredame, 2016; Quinn *et al*., 2017). Proportionality can be thought of as a modified VLR that uses information about the variability of individual components to constrain VLR in a practical range. Proportionality also depends on the reference used for transformation and, unlike SparCC and SPIEC-EASI, does not assume the underlying structure of the associations are sparse (Quinn *et al*., 2018). There are typically three flavours of proportionality including ϕ, ϕ_s and ρ_p (Lovell *et al*., *2015*). The proportionality measures ϕ and ϕ_s both range between [0, ∞), similar to VLR, *as the asymmetric and symmetric versions respectively (*Quinn *et al*., 2017*). The proportionality measure* ρ_p *is the most akin to correlation as the pairwise application results in a symmetric matrix with values ranging from [−1,1] where a value of 1 indicates perfect proportionality amongst components (*Erb and Notredame, 2016*). A major advantage of proportionality measures is that they are robust when analysing relative data (*Lovell *et al*., 2015*) and tend not to produce spurious connections (*Quinn *et al*., 2017*); a stark contrast with Pearson's correlation coefficient which had considerable limitations when applied to compositional data. The properties of robustness to spurious results, scale invariance and interpretability positions proportionality as an effective association metric when inferring cooccurrence (*Bian *et al*., 2017*) and coexpression (*Lovell *et al*., 2015*) from NGS-derived datasets*.

Association measures often involve some type of data transformation in network analysis pipelines to achieve specific weight distributions. It is important to note that many algorithms require weights to be positive real numbers but some algorithms such as the Bellman–Ford shortest path algorithm can handle negative weights (Bang-Jensen and Gutin, 2009). Two common techniques to fulfil the weight assumptions are to investigate: (i) unsigned relationships where only the magnitude of the association is considered ($\boldsymbol{A}_{ij} = |\boldsymbol{\rho}|^{\boldsymbol{\beta}}$; or (ii) signed networks where the weights are forced into the interval [0,1] $\boldsymbol{A}_{ij} = (0.5 + 0.5\ \boldsymbol{\rho})^{\boldsymbol{\beta}}$; where ρ represents the association and β represents the soft thresholding power when $\beta > 1$, or (iii) direction-specific relationships by analysing solely positive associations and masking negative associations (or the reverse). Each of these approaches produce unique network topologies and can be utilized to address different hypotheses depending on whether or not the type of interaction, either positive or negative, is relevant.

Determining network structure and spurious connections

Thresholding and inference are two techniques used to select which connections are represented in a network. Commonly used thresholding methods include: (i) hard thresholding and (ii) soft thresholding. Hard thresholding refers to a binary decision that defines two nodes to be connected (e.g. $\boldsymbol{w} \geq 0.7$). Such thresholds are often arbitrarily chosen without any statistical reasoning and could overlook potentially informative interactions which has been previously investigated (Connor *et al*., 2017). Another approach towards threshold selection is via permutation test to identify significant connections usually followed by multiple tests corrections such as false discovery rate (Unpingco, 2019; Pitman, 1937; Efron and Tibshirani, 1994). Soft thresholding refers to a method where the edge weights are shrunk towards zero and can be applied to networks with edge weights in the range of [0,1] by raising the weight to a power β emphasizing strong connections at the expense of weak connections (Langfelder and Horvath, 2008).

In most networks, the true topology of the network is not known *a priori* and must be inferred via computational methods. Therefore, the concept of false positive edges (i.e. spurious connections) based on true network topology is rather abstract and difficult to assess. In addition to this uncertainty, there exists the Simpson's paradox wherein associations can reverse or disappear when data sets are combined and analysed together (Kievit *et al*., 2013). In an effort to evaluate the presence of false positives, empirically derived functional pathways have been used as ground truth connections (true positives) and compared the associated gene sets using association-based network analysis with inferred structure (Parsana *et al*., 2019). In this paradigm, false positives were defined as edges that were observed between a pair of genes in the inferred network but absent in the list of curated connections. Spurious connections can often occur via the outlier effect (Heyer *et al*., 1999). For example, if the normalized abundance of two features are unrelated in all but a single observation, then the correlation coefficient may be much higher and can result in an inverse relationship (e.g. $\boldsymbol{\rho} = 0.87$ vs. −0.29) (Heyer *et al*., 1999). The topological overlap measure (TOM) is a powerful transformation for symmetric adjacency matrices, particularly for $\boldsymbol{w} \in [0,1]$, that considers pairs of nodes in relation to all nodes within the network instead of in isolation (Yip and Horvath, 2006; Ravasz

et al., 2002). The TOM-based adjacency is particularly useful when the original adjacency matrix is sparse or susceptible to noise by replacing the isolated connections with weighted neighbourhood overlaps, thus, decreasing the effects of spurious or weak connections leading to more robust networks (Yip and Horvath, 2006; Dong and Horvath, 2007; Li and Horvath, 2007; Song *et al*., 2012). Association networks transformed via TOM often reduce or eliminate the number of false positive connections introduced by spurious correlations (Voigt and Almaas, 2019).

Some NGS-derived compositional datasets such as 16/18S amplicon sequencing are often sparse (Paulson *et al*., 2013; Kumar *et al*., 2018; Martino *et al*., 2019) depending on the diversity between samples (e.g. comparing multiple ecosystems or host body sites). Network analysis applied to sparse datasets are subject to false positives when using methods that are not designed for sparsity. *Sparse Correlations for Compositional data* (SparCC) is a technique for inferring correlations from compositional data often used for network analysis with the assumptions that the number of components is large and the true correlation network is sparse (Friedman and Alm, 2012). *Sparse Inverse Covariance Estimation for Ecological Association Inference* (SPIEC-EASI) is another statistical method for the inference of compositional networks, designed for ecologically derived datasets, seeking to address the 'curse-of-dimensionality' with a graphical model inference framework that relies on algorithms for sparse neighbourhood and inverse covariance selection (Kurtz *et al*., 2015). *Regularized estimation of the basis covariance based on compositional data* (REBECCA) identifies significant co-occurrence patterns by finding sparse solutions in a system with a deficient rank and estimating correlations between pairs of basis abundance using log-ratio transformation of counts (Ban *et al*., 2015). The network method used should be determined by the goal of the analysis. For instance, if the aim is to compare differential edges between a treatment system and a reference system, then it may be beneficial to use TOM for fully connected networks that are directly comparable. In contrast, if false positives are a critical concern or community detection is the aim, then it may be more appropriate to use one or a combination of the sparse methods described above.

Interpreting networks and evaluation metrics

Interpreting networks is often the limiting factor for applications beyond visualization. Fortunately, network theory offers several metrics that can be used to describe a particular network at varying levels of abstractions ranging from the network as a whole to objects including nodes and edges. One of the advantages of systems-wide analysis through networks is the ability identify and rank the most *important* nodes, or hubs, in a system (Layeghifard *et al*., 2017). Within a single organism gene expression network, the most important node could be a critical transcription factor, whereas in an environmental system, a hub node could be an organism essential for community stability. The most common metrics for static networks are degree, connectivity, flow and centrality. For simple networks (undirected and unweighted), the degree of a network is the number of connections a particular node contains. In an undirected weighted network, such as the association networks mentioned above, weighted-degree can be computed by summing the weighted connections for each node. The implementation of this weighted-degree is often referred to as connectivity such as in the intramodular connectivity calculations or advanced network visualization methods (Krzywinski *et al*., 2012). It should be noted that the aforementioned description of connectivity as analogous to weighted-degree is informal and the technical definition requires that node connectivity is equal to the minimum number of nodes that must be removed to disconnect the graph (Esfahanian, 2019). In a fully connected undirected network, weighted-degree is an extremely useful metric for measuring connectedness of a node within the network or a sub-network (e.g. intra-genus connectivity in microbial cooccurrence network). An unweighted-degree would yield a uniform distribution because each node is promiscuously connected to the rest in a fully connected network. In directed graphs, a node contains both an in- and out-degree which corresponds to connections into and outwards from a node, respectively. Extending the concepts of directed degree metrics, flow represents the difference between the out- and in-degree with positive and negative measures representing *sources* and *sinks* respectively (Krzywinski *et al*., 2012). In this paradigm, one may utilize Google's *PageRank* (Brin and Page, 1998; Page *et al*., 1999), a powerful link analysis algorithm, which is a variant of eigenvector centrality designed for quantifying the relative importance of a node within a network based on the directed flow of edges into and out of a node. Centrality comes in many different flavours such as eigenvector centrality (a measure of influence within a graph), closeness centrality (the average length of the shortest path between a node and all other nodes in the graph) and betweenness centrality (the frequency in which a node acts as a bridge along the shortest path between two other nodes). As with any analysis, the usefulness of each of these metrics is dependent on the research question, the structure of the graph and complexity of the edges. It is important to note that centrality measures have their caveats in that they underestimate

the power of non-hub nodes due to heterogeneous topology of complex networks (Šikić *et al*., 2013; Layeghifard *et al*., 2017) and do not measure the difference between nodes (Bauer and Lizier, 2012).

Scale-free and heavy-tailed degree distribution topology

A network's organization is characterized by its structure; most notably, the distribution of node degrees. In a random network, the degree distribution is normally distributed. In complex systems, it is common for networks to self-organize into a scale-free state in that the probability $P(k)$ that a node in the network interacts with k other nodes decays as a power law following $P(k) \sim k^{-a}$ (Barabasi and Albert, 1999; Jeong *et al*., 2000; Barabási and Bonabeau, 2003). Scale-free networks are heterogenous, and their topology is dominated by a few highly connected nodes, referred to as hubs, which connect to the rest of the system (Zhang and Horvath, 2005). In network analysis, an examination of hubs typically represents influence within the system such as yeast protein–protein interaction networks and the relevance to proteins essential for survival (Jeong *et al*., 2001; Carter *et al*., 2004; Han *et al*., 2004). It should be noted that the prevalence of scale-free topologies in the natural world have been debated and suggested to be overestimated (Clauset *et al*., 2009; Mitchell, 2009; Broido and Clauset, 2019); therefore, assumptions on scale-free topologies should be properly assessed when exploring various preprocessing metrics, association measures and adjacency transformations.

Recent efforts have described the inconsistencies involving the applications of scale-free topologies with a meta-analysis using a large corpus of published networks, ranging from social networks to biological systems, by characterizing the extent of scale-free topologies including not scale-free, super-weak, weak, strong and strongest (Broido and Clauset, 2019). The findings from *Broido & Clauset et al*. revealed that scale-free structure is not universal, varies across domains and is often confounded as a generic stand-in for other heavy-tailed distributions such as log-normal. Regardless, compared to the entire *Index of Complex Networks* (ICON) corpus (N = 928 networks) with networks from biological, information, social, technological and transportation domains, biological networks were more likely to display the strongest level of direct evidence of scale-free structure.

Scale-free topology is well defined for simple networks (i.e. undirected, unweighted, and monoplex); although, the definition naturally generalizes to weighted networks where k takes on non-negative real numbers (Zhang and Horvath, 2005). In this generalization, scale-free topology can be approximated via the model fitting index R^2 of the linear regression modelled as $\log(p(k)) \sim \log(k)$ (Zhang and Horvath, 2005). Despite being better modelled using an exponentially truncated power law $p(k) \sim k^{-\gamma} \exp(-\alpha k)$ (Csanyi and Szendroi, 2003), *Zhang et al*. suggest the α and γ provide too much flexibility in curve fitting as R^2 is often more robust to adjacency parameters.

Advanced network analysis approaches

Differential networks

Comparing static networks, often referred to as cross-sectional networks, via differential network analysis (DiN [A]) is non-trivial and pertains to interrogating changes in feature interactions (i.e. edges) rather than the changes in the feature measurements (i.e. nodes). In particular, DiNA measures changes in network structure including topological restructuring and edge weights between different states (Lichtblau *et al*., 2016). DiNA is a fusion of two well-studied fields, namely differential abundance analysis and network theory. Differential abundance analysis has been routinely applied to RNA-seq (Robinson *et al*., 2010; Paulson *et al*., 2013; Love *et al*., 2014; Pimentel *et al*., 2017), whereas the latter, network theory, has been a studied for decades (Harary, 1969) with applications in biology to study the centrality of features in a disease network (Joy *et al*., 2005; Wang *et al*., 2011; Winter *et al*., 2012; Espinoza *et al*., 2018). There also exists compositionally aware differential abundance methods such as the *Analysis of Composition of Microbiomes* (ANCOM) which is done by calculating pairwise log ratios between all components and performing a significance test to determine if there is a significant difference in component ratios with respect to sample groupings of interest (Mandal *et al*., 2015); in addition to the aforementioned *ANOVA-Like Differential Expression* (ALDEx2) (Fernandes *et al*., 2013, 2014), the synergy of differential abundance analysis and biological networks obviates the limitations by considering multiple changes that are associated with differences between connectivity states instead of changes in singular features. DiNA algorithms typically compute network metrics for each network individually (e.g. *weighted-degree*) and interrogated via various statistical tests (Espinoza *et al*., 2018) or qualitatively using advanced visualization techniques such as hive plots (Fig. 3) (Krzywinski *et al*., 2012). The utility of DiNA has been validated through several diverse applications from identifying coexpressed genes related to obesity (Fuller *et al*., 2007), key transcriptional regulators associated with cancer that were undetected by expression levels (Carter *et al*., 2004; Lai *et al*., 2004; Choi *et al*., 2005) and regulatory mechanisms in yeast (Hsu *et al*., 2015).

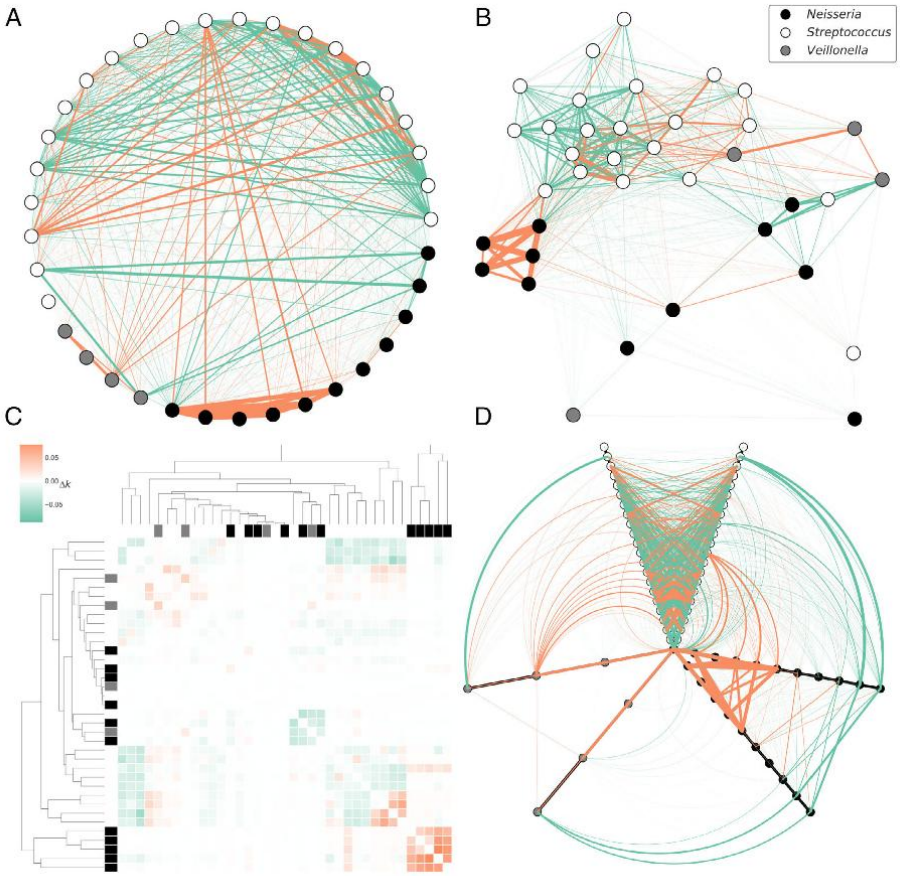


Fig 3. Visualization methods for differential co-occurrence networks.Supragingival plaque oral microbiome 16S community (Gomez *et al*., 2017). Turquoise and orange represent connections that are enriched in healthy and diseased cohorts respectively. Undirected weighted networks were constructed by the following pipeline: (1) removing OTUs that were not present in at least 25 individuals (5% of the cohort); (2) subsetting diseased and healthy cohorts; (3) ρ_p pairwise proportionality of each cohort (Erb and Notredame, 2016; Quinn *et al*., 2017, *compositional*, Espinoza, 2020); (4) signed transformation; (5) soft thresholding ($\beta = 12$); (6) topological overlap measure; (7) DiN = $TOM_{Diseased} - TOM_{Healthy}$; (8) subgraph for top three abundant genera and (9) visualization of (A, B) traditional networks (*NetworkX* (Hagberg *et al*., 2008), (C) clustermaps (*Seaborn*, Waskom *et al*., 2020) and (D) hive plots (*Soothsayer*, Espinoza, 2019).

Temporal networks

Understanding a system in a particular state is useful for some research questions such as investigating the differences between a disease and a non-diseased states (Gomez *et al*., 2017; Espinoza *et al*., 2018). However, it is difficult to address hypotheses about how a system changes over time using static networks, which are based on many assumptions including the following: (1) network topology is fixed; (2) processes of flow are at a steady state and (3) edges represent persistent interactions (Blonder *et al*., 2012). Such static networks can be further extended to study dynamically changing systems such as temporal networks. Network dynamics can be captured by studying the changes in the overall topology, node connectedness, node interactions (i.e. edges) and strength of the interactions (i.e. edge weights). Temporal networks are able to address such network dynamics by extending the edge domain across a time dimension instead of within the confines of a particular cross-sectional timepoint. Although concepts derived from static network theory apply to individual time states within temporal networks, extended theory must be utilized to study the dynamics between states across

time. There exists many network measurements designed for temporal dynamics with the most notable including: *temporal centrality*, *closeness centrality, volatility* and *reachability latency* with many more described in (Thompson *et al*., 2017). Of the centrality based adaptive measures, *temporal centrality* measures the number of overall connections in a timepoint, whereas *closeness centrality* measures the time between specific connections. *Volatility* measures the rate of change in network states with respect to timepoint. Lastly, *reachability latency* is a useful index as it measures the time taken for all nodes to reach each other. Although temporal network theory offers an arsenal of metrics, usage should be crafted towards interpretation of a hypothesis on how states evolve over time. For example, the usage of *volatility* would have a much different interpretation when applied to a temporal resolution of milliseconds compared to staggered years or subjective timepoints (Thompson *et al*., 2017).

There are two main approaches in implementing temporal networks; namely (1) time-ordered where each edge is present only for a precise period and (2) time-aggregated networks over relatively short time frames (Blonder *et al*., 2012; Farine, 2018). Successful applications of temporal network theory have been demonstrated using functional magnetic resonance imaging data to explore dynamic properties of neural connectivity (Thompson *et al*., 2017), drought-responsive plant genes based on differential rhythmic expression profiles (Greenham *et al*., 2017), predicting parasite transmission spread in wild lemurs (Springer *et al*., 2017), recurrent patterns of microdiversity in a temperate coastal marine environments (Chafee *et al*., 2018) and longitudinal multi-omics to advance precision health (Schüssler-Fiorenza Rose *et al*., 2019; Zhou *et al*., 2019). When deciding whether to pursue temporal network analysis or adapt a static framework to incorporate temporal dynamics, it is vital to develop the framework around a particular research question such as available time resolution for samples, whether or not the time-ordered or time-aggregate would better model the hypothesis, or, most importantly, if a temporal component would yield any insight in the base hypothesis.

Sample-specific networks

Sample-specific networks (SSN) have been employed in the quest for personalized medicine to identify patient-specific biomarkers and changes in connectivity that can describe or predict health status (Liu *et al*., 2016; Yu *et al*., 2017; Kuijjer *et al*., 2019). Although previous research has focused on location-specific (Lima-Mendez *et al*., 2015) or host-specific (Ofaim *et al*., 2017) networks, SSNs have shown to be reliable for accurately characterizing a specific disease state within an individual (Liu *et al*., 2016); although, these methods could seamlessly be adapted to investigate environmental systems such as the Tara oceans (Karsenti *et al*., 2011) or Global ocean sampling expedition (Venter *et al*., 2004; Rusch *et al*., 2007) datasets. In a clinical setting, patient-specific diseases generally result from the dysfunction of the underlying system rather than individual molecules (Barabási *et al*., 2011; Hood and Flores, 2012). The assumptions of SSNs posit that if a single sample can significantly alter the association of two features, then the query sample is considered to be inconsistent with the interactions in the reference network. *Liu et al*. developed a statistical method to construct SSNs based on statistical perturbation analysis of a single sample against a group of control samples validated with nine different cancer types from *The Cancer Genome Atlas* dataset (Weinstein *et al*., 2013). In short, this perturbation method computes the pairwise Pearson correlation of a background network compared to the same network with the sample of interest added. *Liu et al*. discovered network patterns unique to specific types of cancer; personalized features revealed important regulatory patterns of driver genes, strong associations between SSNs and somatic mutations and the ability to predict driver genes from edges involving hub genes. The SSN method employed in *Liu et al*. was shown to be less sensitive to spurious associations than standard Pearson correlation-based networks through the incorporation of a reference network for comparisons. SSNs are in their infancy; but, their potential is unprecedented especially when studying how a particular sample state perturbs a reference state. Furthermore, SSNs also present an opportunity to incorporate sample-specific phenotypic data such as nutrient measurements in ecological samples or clinical measurements in medical samples.

Visualizing network complexity

.Network visualizations are infamously difficult to interpret. Visualization is a qualitative assessment of network structure that can guide quantitative assessments downstream. The complexity of a network scales exponentially which can make visualization of large networks difficult. For instance, a fully connected undirected network (e.g. ρ_p proportionality network with soft thresholding) contains $\boldsymbol{(N_n^2 - N_n)/2}$ non-redundant edges ($\mathbf{N_e}$) where $\mathbf{N_n}$ represents the number of nodes. This exponential scale makes visualizing a fully connected undirected network with $N_n = 100$ nodes ($N_e = 4950$) much easier than a network with $N_n = 1000$ nodes ($N_e = 499{,}500$). To demonstrate the interpretability of network visualization methods, we construct a cooccurrence DiN using the top three abundant genera (*Streptococcus*, *Veillonella* and *Neisseria*) from a supragingival plaque oral microbiome dataset (Gomez *et al*., 2017) consisting of subjects with ($N_{diseased} = 126$) and without ($N_{healthy} = 337$) dental caries (Fig. 3). A traditional network visualization would

be the aforementioned 'ball-and-stick' method with efficacy governed by the layout algorithm for positioning nodes in relation to each other in a plot; usually with respect to edge weight. A common network is the circular layout using custom node ordering which is feasible for simple networks but decreases in interpretability with increasing network complexity (Fig. 3A). Networks visualized with layout algorithms that incorporate edge weights are more useful for qualitative visual clustering of nodes such as the *ForceAtlas2* (Jacomy *et al*., 2014), *Fruchterman-Reingold force-directed algorithm* (Fruchterman and Reingold, 1991) or *graphviz* algorithms (Ellson *et al*., 2001) (Fig. 3B). However, this method of visualization is often only interpretable for fully connected networks with low complexity or scale-free networks and can quickly turn into the notorious 'hairball' network visualization. Despite the archetype of visualizing networks with layout-based methods, they are often difficult to interpret because their creation is often driven by an aesthetic heuristic which influences how the topology is rendered (Krzywinski *et al*., 2012). More complex networks benefit more from alternative visualization approaches such as clustermaps and hive plots. Clustermaps are a unique combination of agglomerative hierarchical clustering and matrix heatmaps leveraging unsupervised relationships between nodes and the visualization of edge weights as values within the heatmap (Fig. 3C). Clustermaps are beneficial for visualizing low and medium complexity networks but can be uninterpretable, difficult to render and computationally expensive to compute as the number of nodes increases. Hive plots are a network visualization method that are applicable at all levels of complexity (Krzywinski *et al*., 2012). The hive plot places nodes on a radially oriented linear axis with nodes positioned along the axis either by structural properties or user-defined selections. Hive plots contain three flexible components including: (1) the assignment of node coordinates to and within an axis; (2) the layout profile of each axes (position, scale and angle) and (3) the aesthetics of (weighted) edges visualized as curves between nodes for intra/inter-axes connectivity. The hive plot most effectively illustrates enriched intra-genus connectivity within *Neisseria* in the diseased cohort and phenotype-specific inter-genus connectivity profiles between specific OTUs (Fig. 3D). Hive plots are powerful for networks but are difficult to adapt for temporal networks. Arguably, the most intuitive temporal network is the slice plot implemented through the *teneto* Python package (Thompson *et al*., 2017). A slice plot essentially decomposes a hive into a collection of arc plots, a linear segment with nodes positioned along the line and arcs showing connections, juxtaposed consecutively to visualize connections that remain consistent or, alternatively, inconsistent with respect to time. In networks with few connections (e.g. $N_n \leq 100$), the labels can be useful in discerning connections, but larger networks are often assessed globally without visualizing the label of nodes directly. Network visualization methods are still evolving and can benefit from the insight of researchers with unique hypotheses and computational abilities.

Summary and outlook

NGS-derived datasets are compositional and should be considered as compositions at all stages of analysis (Gloor *et al*., 2017). We have overviewed the characteristics of compositional data, the bias that occur when compositional datasets are analysed inappropriately and transformation techniques that mitigate the bias such as log-ratio transformations. In addition, for compositional data analysis, we detailed the advantages and caveats of various methods to construct association networks. We further reviewed the analytical metrics for quantifying different aspects of network topology and application of advanced network analysis to model more complex systems.

Despite the progressive techniques recently developed to interpret biological systems, the nascent field of systems biology is far from the status of omniscient. Not knowing the true topology of a system *a priori* inherently limits our approaches towards fully understanding a system's natural complexity. Furthermore, biological systems are not static and modelling the transition between states will yield more intuitive insights on the schematics of these complex structures. The aphorism that 'all models are wrong, but some are useful' (Box *et al*., 2009) holds truth in the paradigm of inference-based systems biology where knowing the true network structure of an abstract space *a priori* is not attainable. Biological systems are complex because they are abstract constructs used to model an observed phenomenon. This complexity is the aftermath of the uncertainty of true associations, the sensitivity of the methods to infer associations, unaccounted variance (e.g. unknown phenotype) and the dynamics of how these abstractions, such edge weights and node inclusion, evolve over time. The abstract space defined by a network is the source of its versatility while also representing the crux of germane interpretation.

The advanced network approaches described in this review, combinations thereof, or even *networks-of-networks* (Gao *et al*., 2014) can potentially be utilized to address humanity's most pressing issues. For instance, consider the topic of drug discovery in the scope of antibiotic-resistance. Imagine one has identified a novel chemical entity (NCE) from a soil microbiome, referred to as NCE_X, that appears to exhibit a unique mechanism of action against a particular pathogenic organism (*P. organism*). One may ask how the query NCE perturbs the baseline state ($\boldsymbol{t_0}$) of *P. organism* with dose $\boldsymbol{d}$ at different time intervals? This question could be addressed by creating a differential network at each timepoint $\mathbf{n}$ ($DiNET_n$) as $\boldsymbol{t_0}$ vs. $\boldsymbol{t_n}$ where each static network is derived via coexpression-based topological overlap

measures. Of greater insight, albeit greater complexity, one may ask how the perturbations of NCE_x-challenged *P. organism* at dose ***d*** compares to a negative control, challenged with a solvent such as water, over a time interval from 0 min to 1 h? The symbiosis of temporal, sample-specific and differential correlation networks could be used to investigate this question which, naturally, could guide the experimental design for the over-arching project. SSNs have already been harnessed for personalized medicine and could be further augmented by incorporating not only temporal dimensionality but multimodality. For more grandiose applications, imagine the synergy of explainable artificial intelligence (Gunning, 2017), system-wide cellular modelling (Ebrahim *et al*., 2013) and 'network-of-networks' (multi-level network) frameworks (Gao *et al*., 2014) harnessed by domain experts spanning climate science to microbiology, public health to agriculture and from economics to politics modelling the complex flux of resources; an interdisciplinary effort to usurp climate change by identifying solution states that are not only environmentally sustainable but economically productive.

The future of systems biology must be approached from creative vantage points by building combinatorically on the cornerstones of established concepts, understanding the assumptions of various statistical methods and interpreting these mathematical abstractions in the context of insightful biological questions where domain knowledge is of utmost importance. The synergy of domain expertise, advanced analytical methods and creative minds is the foundation of cutting-edge science. Modelling complex systems has provided insight in the past and will certainly continue to do so in the future with the evolution of network theory, and the inventiveness catalysed by the human mind and machines to decipher latent patterns embedded within natural and abstract systems.

Conflict of interest

None.

Acknowledgements

This work was supported by NSF grant OCE-1558453 (to CLD), NASA NNA15BB034A (to CLD) and P01AI118687 (to KEN).

References

Aitchison, J. (1982) The statistical analysis of compositional data. *J R Stat Soc Ser B* **44**: 139–160.

Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J.A., and Pawlowsky-Glahn, V. (2000) Logratio analysis and compositional distance. *Math Geol* **32**: 271–275.

Amir, A., McDonald, D., Navas-Molina, J.A., Kopylova, E., Morton, J.T., Zech Xu, Z., *et al*. (2017) Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems* **2**: e00191-16. https://msystems.asm.org/content/2/2/e00191-16.

Ban, Y., An, L., and Jiang, H. (2015) Investigating microbial co-occurrence patterns based on metagenomic compositional data. *Bioinformatics* **31**: 3322–3329.

Bang-Jensen, J., and Gutin, G. (2009) *Digraphs: Theory, Algorithms, and Applications*, London, England: Springer-Verlag.

Barabasi, A.-L., and Albert, R. (1999) Emergence of scaling in random networks. *Science* **286**: 509–512.

Barabási, A.-L., and Bonabeau, E. (2003) Scale-free networks. *Sci Am* **288**: 60–69.

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011) Network medicine: a network-based approach to human disease. *Nat Rev Genet* **12**: 56–68.

Bauer, F., and Lizier, J.T. (2012) Identifying influential spreaders and efficiently estimating infection numbers in epidemic models: a walk counting approach. *EPL Europhys Lett* **99**: 68007.

Bellman, R. (2003) *Dynamic Programming*, United States: Dover Publications.

Bian, G., Gloor, G.B., Gong, A., Jia, C., Zhang, W., Hu, J., Zhang, H., Zhang, Y., Zhou, Z., Zhang, J., Burton, J.P., Reid, G., Xiao, Y., Zeng, Q., Yang, K., and Li, J. (2017). The gut microbiota of healthy aged chinese is similar to that of the healthy young. *mSphere* **2**. http://dx.doi.org/10.1128/msphere.00327-17.

Bishop, C.M. (2006) *Pattern Recognition and Machine Learning*, New York, NY: Springer-Verlag.

Blasco-Moreno, A., Pérez-Casany, M., Puig, P., Morante, M., and Castells, E. (2019) What does a zero mean? Understanding false, random and structural zeros in ecology. *Methods Ecol Evol* **10**: 949–959.

Blonder, B., Wey, T.W., Dornhaus, A., James, R., and Sih, A. (2012) Temporal dynamics and network analysis. *Methods Ecol Evol* **3**: 958–972.

Bokulich, N.A., Rideout, J.R., Mercurio, W.G., Shiffer, A., Wolfe, B., Maurice, C.F., *et al*. (2016) Mockrobiota: a public resource for microbiome bioinformatics benchmarking. *mSystems* **1**: e00062-16.

Box, G.E.P., Luceño, A., and Paniagua-Quiñones, M.D.C. (2009) *Statistical Control by Monitoring and Adjustment*. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Brin, S. and Page, L. (1998). *The Anatomy of a Large-Scale Hypertextual Web Search Engine* (107–117). Amsterdam, Netherlands: WWW7: Proceedings of the seventh international conference on World Wide Web 7.

Broido, A.D., and Clauset, A. (2019) Scale-free networks are rare. *Nat Commun* **10**: 1017.

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., and Holmes, S.P. (2016) DADA2: high-resolution sample inference from Illumina amplicon data. *Nat Methods* **13**: 581–583.

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Lozupone, C.A., Turnbaugh, P.J., *et al*. (2011) Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci U S A* **108**: 4516–4522.

Carter, S.L., Brechbuhler, C.M., Griffin, M., and Bond, A.T. (2004) Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics* **20**: 2242–2250.

Chafee, M., Fernàndez-Guerra, A., Buttigieg, P.L., Gerdts, G., Eren, A.M., Teeling, H., and Amann, R.I. (2018) Recurrent patterns of microdiversity in a temperate coastal marine environment. *ISME J* **12**: 237–252.

Checinska Sielaff, A., Urbaniak, C., Mohan, G.B.M., Stepanov, V.G., Tran, Q., Wood, J.M., *et al*. (2019) Characterization of the total and viable bacterial and fungal communities associated with the international Space Station surfaces. *Microbiome* **7**: 50–50.

Chen, J., and Li, H. (2013) Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *Ann Appl Stat* **7**: 418–442.

Choi, J.K., Yu, U., Yoo, O.J., and Kim, S. (2005) Differential coexpression analysis using microarray data and its application to human cancer. *Bioinformatics* **21**: 4348–4355.

Clauset, A., Shalizi, C.R., and Newman, M.E.J. (2009) Power-law distributions in empirical data. *SIAM Rev* **51**: 661–703.

Connor, N., Barberán, A., and Clauset, A. (2017) Using null models to infer microbial co-occurrence networks. *PLoS One* **12**: e0176751.

Crits-Christoph, A., Diamond, S., Butterfield, C.N., Thomas, B.C., and Banfield, J.F. (2018) Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis. *Nature* **558**: 440–444.

Csányi, G., and Szendrői, B. (2004). Structure of a large social network. *Physical Review E* **69**. http://dx.doi.org/10.1103/physreve.69.036131.

Diaconis, P., Goel, S., and Holmes, S. (2008) Horseshoes in multidimensional scaling and local kernel methods. *Ann Appl Stat* **2**: 777–807.

Dong, J., and Horvath, S. (2007) Understanding network concepts in modules. *BMC Syst Biol* **1**: 24.

Ebrahim, A., Lerman, J.A., Palsson, B.O., and Hyduke, D.R. (2013) COBRApy: COnstraints-based reconstruction and analysis for python. *BMC Syst Biol* **7**: 74.

Edgar, R.C. (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Methods* **10**: 996–998.

Efron, B., and Tibshirani, R. (1994) *An Introduction to the Bootstrap*, London, England: Chapman & Hall.

Egozcue, J.J., and Pawlowsky-Glahn, V. (2005) Groups of parts and their balances in compositional data analysis. *Math Geol* **37**: 795–828.

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barceló-Vidal, C. (2003) Isometric logratio transformations for compositional data analysis. *Math Geol* **35**: 279–300.

Ellson, J., Ellson, J., Gansner, E., Koutsofios, L., North, S., Woodhull, G., *et al*. (2001) Graphviz — open source graph drawing tools. *Lect NOTES Comput Sci* **2265**: 483–484.

Erb, I., and Notredame, C. (2016) How should we measure proportionality on relative gene expression data? *Theory Biosci* **135**: 21–36.

Esfahanian, A.-H. (2019). *Connectivity Algorithms*, Michigan, MI: Michigan State University. http://www.cse.msu.edu/∼cse835/Papers/Graph_connectivity_revised.pdf.

Espinoza, J.L. (2019) *soothsayer: High-level package for (Bio-)informatics*. GitHub. https://github.com/jolespin/soothsayer.

Espinoza, J.L., Harkins, D.M., Torralba, M., Gomez, A., Highlander, S.K., Jones, M.B., *et al*. (2018) Supragingival plaque microbiome ecology and functional potential in the context of health and disease. *MBio* **9**: e01631-18.

Fang, H., Huang, C., Zhao, H., and Deng, M. (2015) CCLasso: correlation inference for compositional data through Lasso. *Bioinformatics* **31**: 3172–3180.

Farine, D.R. (2018) When to choose dynamic vs. static social network analysis. *J Anim Ecol* **87**: 128–138.

Faust, K., and Raes, J. (2016) CoNet app: inference of biological association networks using Cytoscape. *F1000Research* **5**: 1519.

Fernandes, A.D., Macklaim, J.M., Linn, T.G., Reid, G., and Gloor, G.B. (2013) ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-Seq. *PLoS One* **8**: e67019.

Fernandes, A.D., Reid, J.N.S., Macklaim, J.M., McMurrough, T.A., Edgell, D.R., and Gloor, G.B. (2014) Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome* **2**: 15.

Finucane, M.M., Sharpton, T.J., Laurent, T.J., and Pollard, K.S. (2014) A taxonomic signature of obesity in the microbiome? Getting to the guts of the matter. *PLoS One* **9**: e84689.

Friedman, J., and Alm, E.J. (2012) Inferring correlation networks from genomic survey data. *PLoS Comput Biol* **8**: e1002687.

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

Fruchterman, T.M.J., and Reingold, E.M. (1991) Graph drawing by force-directed placement. *Softw Pract Exp* **21**: 1129–1164.

Fuller, T.F., Ghazalpour, A., Aten, J.E., Drake, T.A., Lusis, A.J., and Horvath, S. (2007) Weighted gene coexpression network analysis strategies applied to mouse weight. *Mamm Genome* **18**: 463–472.

Gao, J., Li, D., and Havlin, S. (2014) From a single network to a network of networks. *Natl Sci Rev* **1**: 346–356.

Gloor, G.B., Macklaim, J.M., Pawlowsky-Glahn, V., and Egozcue, J.J. (2017) Microbiome datasets are compositional: and this is not optional. *Front Microbiol* **8**: 2224.

Goh, W.W.B., Wang, W., and Wong, L. (2017) Why batch effects matter in omics data, and how to avoid them. *Trends Biotechnol* **35**: 498–507.

Gomez, A., Espinoza, J.L., Harkins, D.M., Leong, P., Saffery, R., Bockmann, M., *et al*. (2017) Host genetic control of the oral microbiome in health and disease. *Cell Host Microbe* **22**: 269-278.e3.

Goodrich, J.K., Davenport, E.R., Beaumont, M., Jackson, M. A., Knight, R., Ober, C., *et al*. (2016) Genetic determinants of the gut microbiome in UK twins. *Cell Host Microbe* **19**: 731–743.

Greenham, K., Guadagno, C.R., Gehan, M.A., Mockler, T.C., Weinig, C., Ewers, B.E., and McClung, C.R. (2017). Temporal network analysis identifies early physiological and

transcriptomic indicators of mild drought in Brassica rapa. *eLife* **6**. http://dx.doi.org/10.7554/elife.29655.

Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlimi, A., Roux, S., *et al*. (2016) Plankton networks driving carbon export in the oligotrophic ocean. *Nature* **532**: 465–470.

Gunning, D. (2017) *Explainable Artificial Intelligence (XAI)*, Arlington County, VA: DARPA. https://www.darpa.mil/attachments/XAIProgramUpdate.pdf.

Hagberg, A.A., Schult, D.A., and Swart, P.J. (2008) *Exploring Network Structure, Dynamics, and Function using NetworkX*, Pasadena, CA: Proceedings of the 7th Python in Science Conference. https://www.osti.gov/biblio/960616-exploring-network-structure-dynamics-function-using-networkx.

Han, J.-D.J., Bertin, N., Hao, T., Goldberg, D.S., Berriz, G.F., Zhang, L.V., *et al*. (2004) Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature* **430**: 88–93.

Harary, F. (1969) *Graph Theory*, New York, NY: Avalon Publishing.

Hardin, J., Mitani, A., Hicks, L., and VanKoten, B. (2007) A robust measure of correlation between two genes on a microarray. *BMC Bioinform* **8**: 220.

Harrison, J.G., Calder, W.J., Shastry, V., and Buerkle, C.A. (2019) Dirichlet-multinomial modelling outperforms alternatives for analysis of microbiome and other ecological count data. *Mol Ecol Resour* **20**: 481–497.

Heyer, L.J., Kruglyak, S., and Yooseph, S. (1999) Exploring expression data: identification and analysis of coexpressed genes. *Genome Res* **9**: 1106–1115.

Holmes, I., Harris, K., and Quince, C. (2012) Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLoS One* **7**: e30126.

Hood, L., and Flores, M. (2012) A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory. *N Biotechnol* **29**: 613–624.

Hsu, C.-L., Juan, H.-F., and Huang, H.-C. (2015) Functional analysis and characterization of differential coexpression networks. *Sci Rep* **5**: 13295.

Jacomy, M., Venturini, T., Heymann, S., and Bastian, M. (2014) ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS One* **9**: e98679.

Jeong, H., Mason, S.P., Barabási, A.-L., and Oltvai, Z.N. (2001) Lethality and centrality in protein networks. *Nature* **411**: 41–42.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., and Barabási, A.-L. (2000) The large-scale organization of metabolic networks. *Nature* **407**: 651–654.

Joy, M.P., Brock, A., Ingber, D.E., and Huang, S. (2005) High-betweenness proteins in the yeast protein interaction network. *J Biomed Biotechnol* **2005**: 96–103.

Karsenti, E., Acinas, S.G., Bork, P., Bowler, C., De Vargas, C., Raes, J., *et al*. (2011) A holistic approach to marine eco-systems biology. *PLoS Biol* **9**: e1001177.

Kievit, R.A., Frankenhuis, W.E., Waldorp, L.J., and Borsboom, D. (2013) Simpson's paradox in psychological science: a practical guide. *Front Psychol* **4**: 513.

Knights, D., Costello, E.K., and Knight, R. (2011) Supervised classification of human microbiota. *FEMS Microbiol Rev* **35**: 343–359.

Krzywinski, M., Birol, I., Jones, S.J., and Marra, M.A. (2012) Hive plots–rational approach to visualizing networks. *Brief Bioinform* **13**: 627–644.

Kuhnert, P.M., Martin, T.G., Mengersen, K., and Possingham, H.P. (2005) Assessing the impacts of grazing levels on bird density in woodland habitat: a Bayesian approach using expert opinion. *Environ* **16**: 717–747.

Kuijjer, M.L., Tung, M.G., Yuan, G., Quackenbush, J., and Glass, K. (2019) Estimating sample-specific regulatory networks. *iScience* **14**: 226–240.

Kumar, M.S., Slud, E.V., Okrah, K., Hicks, S.C., Hannenhalli, S., and Corrada Bravo, H. (2018) Analysis and correction of compositional bias in sparse sequencing count data. *BMC Genom* **19**: 799.

Kurtz, Z.D., Müller, C.L., Miraldi, E.R., Littman, D.R., Blaser, M.J., and Bonneau, R.A. (2015) Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput Biol* **11**: e1004226.

Lachmann, A., Giorgi, F.M., Lopez, G., and Califano, A. (2016) ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics* **32**: 2233–2235.

Laczny, C.C., Sternal, T., Plugaru, V., Gawron, P., Atashpendar, A., Margossian, H., *et al*. (2015) VizBin - an application for reference-independent visualization and human-augmented binning of metagenomic data. *Microbiome* **3**: 1.

Lai, Y., Wu, B., Chen, L., and Zhao, H. (2004) A statistical method for identifying differential gene-gene co-expression patterns. *Bioinformatics* **20**: 3146–3155.

Langfelder, P., and Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform* **9**: 559.

Langfelder, P., and Horvath, S. (2012) Fast R functions for robust correlations and hierarchical clustering. *Journal of statistical software* **46**.

Layeghifard, M., Hwang, D.M., and Guttman, D.S. (2017) Disentangling interactions in the microbiome: a network perspective. *Trends Microbiol* **25**: 217–228.

Levering, J., Dupont, C.L., Allen, A.E., Palsson, B.O., and Zengler, K. (2017) Integrated regulatory and metabolic networks of the marine diatom *Phaeodactylum tricornutum* predict the response to rising CO2 levels. *mSystems* **2**: e00142-16.

Li, A., and Horvath, S. (2007) Network neighborhood analysis with the multi-node topological overlap measure. *Bioinformatics* **23**: 222–231.

Lichtblau, Y., Zimmermann, K., Haldemann, B., Lenze, D., Hummel, M., and Leser, U. (2016) Comparative assessment of differential network analysis methods. *Brief Bioinform* **18**: bbw061.

Lima-Mendez, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F., *et al*. (2015) Determinants of community structure in the global plankton interactome. *Science* **348**: 1262073.

Ling, L.L., Schneider, T., Peoples, A.J., Spoering, A.L., Engels, I., Conlon, B.P., *et al*. (2015) A new antibiotic kills pathogens without detectable resistance. *Nature* **517**: 455–459.

Liu, X., Wang, Y., Ji, H., Aihara, K., and Chen, L. (2016) Personalized characterization of diseases using sample-specific networks. *Nucleic Acids Res* **44**: e164.

Logares, R., Deutschmann, I.M., Junger, P.C., Giner, C.R., Krabberød, A.K., Schmidt, T.S.B., *et al*. (2020) Disentangling the mechanisms shaping the surface ocean microbiota. *Microbiome* **8**: 55.

Love, M.I., Huber, W., and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550.

Lovell, D., Pawlowsky-Glahn, V., Egozcue, J.J., Marguerat, S., and Bähler, J. (2015) Proportionality: a valid alternative to correlation for relative data. *PLoS Comput Biol* **11**: e1004075.

Mandakovic, D., Rojas, C., Maldonado, J., Latorre, M., Travisany, D., Delage, E., *et al*. (2018) Structure and co-occurrence patterns in microbial communities under acute environmental stress reveal ecological factors fostering resilience. *Sci Rep* **8**: 1–12.

Mandal, S., Van Treuren, W., White, R.A., Eggesbø, M., Knight, R., and Peddada, S.D. (2015) Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial Ecology in Health & Disease* **26**: 27663. https://doi,org/10.3402/mehd.v26.27663.

Martin, T.G., Wintle, B.A., Rhodes, J.R., Kuhnert, P.M., Field, S.A., Low-Choy, S.J., *et al*. (2005) Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. *Ecol Lett* **8**: 1235–1246.

Martino, C., Morton, J.T., Marotz, C.A., Thompson, L.R., Tripathi, A., Knight, R., and Zengler, K. (2019) A novel sparse compositional technique reveals microbial perturbations. *mSystems* **4**: e00016-19.

Mcgeachie, M.J., Sordillo, J.E., Gibson, T., Weinstock, G.M., Liu, Y.Y., Gold, D.R., *et al*. (2016) Longitudinal prediction of the infant gut microbiome with dynamic Bayesian networks. *Sci Rep* **6**: 1–11.

McInnes, L., Healy, J., and Melville, J. (2018) *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, Ithaca, NY: arXiv. http://arxiv.org/abs/1802.03426.

Mitchell, M. (2009) *Complexity: A Guided Tour*. New York, NY, USA: Oxford University Press, Inc.

Morton, J.T., Marotz, C., Washburne, A., Silverman, J., Zaramela, L.S., Edlund, A., *et al*. (2019) Establishing microbial composition measurement standards with reference frames. *Nat Commun* **10**: 1–11.

Morton, J.T., Sanders, J., Quinn, R.A., McDonald, D., Gonzalez, A., Vázquez-Baeza, Y., *et al*. (2017) Balance trees reveal microbial niche differentiation. *mSystems* **2**: e00162-16.

Morton, J.T., Toran, L., Edlund, A., Metcalf, J.L., Lauber, C., and Knight, R. (2017) Uncovering the horseshoe effect in microbial analyses. *mSystems* **2**: e00166-16.

Ofaim, S., Ofek-Lalzar, M., Sela, N., Jinag, J., Kashi, Y., Minz, D., and Freilich, S. (2017) Analysis of microbial functions in the rhizosphere using a metabolic-network based framework for metagenomics interpretation. *Front Microbiol* **8**: 1606.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999) *The PageRank Citation Ranking: Bringing Order to the Web*, Stanford, CA: Stanford. https://www.semanticscholar.org/paper/The-PageRank-Citation-Ranking%3A-Bringing-Order-to-Page-Brin/eb82d3035849cd23578096462ba419b53198a556.

Parada, A.E., Needham, D.M., and Fuhrman, J.A. (2016) Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environ Microbiol* **18**: 1403–1414.

Parsana, P., Ruberman, C., Jaffe, A.E., Schatz, M.C., Battle, A., and Leek, J.T. (2019) Addressing confounding artifacts in reconstruction of gene co-expression networks. *Genome Biol* **20**: 94.

Paulson, J.N., Stine, O.C., Bravo, H.C., and Pop, M. (2013) Differential abundance analysis for microbial marker-gene surveys. *Nat Methods* **10**: 1200–1202.

Pawlowsky-Glahn, V., and Egozcue, J.J. (2011) Exploring compositional data with the CoDa-Dendrogram. *Austrian J Stat* **40**: 103–113.

Pawlowsky-Glahn, V., Egozcue, J.J., and Tolosana-Delgado, R. (2011) *Lecture Notes on Compositional Data Analysis*, Spain: University of Girona. http://www.compositionaldata.com/material/others/Lecture_notes_11.pdf.

Pawlowsky-Glahn, V., Egozcue, J., and Tolosana-Delgado, R. (2015) *Modeling and Analysis of Compositional Data*, New York, NY: John Wiley & Sons. https://doi.org/10.1002/9781119003144.

Pearson, K. (1896). Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. *Philos Trans R Soc London Ser A, Contain Pap a Math or Phys Character* **187**: 253–318.

Pimentel, H., Bray, N.L., Puente, S., Melsted, P., and Pachter, L. (2017) Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat Methods* **14**: 687–690.

Pine, P.S., Munro, S.A., Parsons, J.R., McDaniel, J., Lucas, A.B., Lozach, J., *et al*. (2016) Evaluation of the external RNA controls consortium (ERCC) reference material using a modified Latin square design. *BMC Biotechnol* **16**: 54.

Pitman, E.J.G. (1937) Significance tests which may be applied to samples from any populations. *Suppl to J R Stat Soc* **4**: 119.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**: 904–909.

Quinn, T.P., Crowley, T.M., and Richardson, M.F. (2018). Benchmarking differential expression analysis tools for RNA-Seq: normalization-based vs. log-ratio transformation-based methods. *BMC Bioinformatics* **19**. http://dx.doi.org/10.1186/s12859-018-2261-8.

Quinn, T.P., Erb, I., Richardson, M.F., and Crowley, T.M. (2018) Understanding sequencing data as compositions: an outlook and review. *Bioinformatics* **34**: 2870–2878.

Quinn, T.P., Richardson, M.F., Lovell, D., and Crowley, T.M. (2017) Propr: an R-package for identifying proportionally abundant features using compositional data analysis. *Sci Rep* **7**: 1–9.

Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., and Barabási, A.-L. (2002) Hierarchical organization of modularity in metabolic networks. *Science* **297**: 1551–1555.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.-F., *et al*. (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**: 431–437.

Rivera-Pinto, J., Egozcue, J.J., Pawlowsky-Glahn, V., Paredes, R., Noguera-Julian, M., and Calle, M. L. (2018). Balances: a new perspective for microbiome analysis. *mSystems* **3**. http://dx.doi.org/10.1128/msystems.00053-18.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010) edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139–140.

Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S., *et al*. (2007) The sorcerer II Global Ocean sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLoS Biol* **5**: 0398–0431.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., *et al*. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**: 7537–7541.

Schüssler-Fiorenza Rose, S.M., Contrepois, K., Moneghetti, K.J., Zhou, W., Mishra, T., Mataraso, S., *et al*. (2019) A longitudinal big data approach for precision health. *Nat Med* **25**: 792–804.

Shomorony, I., Cirulli, E.T., Huang, L., Napier, L.A., Heister, R.R., Hicks, M., *et al*. (2020) An unsupervised learning approach to identify novel signatures of health and disease from multimodal data. *Genome Med* **12**: 7.

Šikić, M., Lančić, A., Antulov-Fantulin, N., and Štefančić, H. (2013) Epidemic centrality — is there an underestimated epidemic impact of network peripheral nodes? *Eur Phys J B* **86**: 1–13.

Silverman, J.D., Washburne, A.D., Mukherjee, S., and David, L.A. (2017). A phylogenetic transform enhances analysis of compositional microbiota data. *eLife* **6**. http://dx.doi.org/10.7554/elife.21887.

Sneath, P.H.A., and Sokal, R.R. (1962) Numerical taxonomy. *Nature* **193**: 855–860.

Song, L., Langfelder, P., and Horvath, S. (2012) Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinform* **13**: 328.

Springer, A., Kappeler, P.M., and Nunn, C.L. (2017) Dynamic vs. static social networks in models of parasite transmission: predicting Cryptosporidium spread in wild lemurs. *J Anim Ecol* **86**: 419–433.

Tackmann, J., Matias Rodrigues, J.F., and von Mering, C. (2019) Rapid inference of direct interactions in large-scale ecological networks from heterogeneous microbial sequencing data. *Cell Syst* **9**: 286-296.e8.

Thompson, W.H., Brantefors, P., and Fransson, P. (2017) From static to temporal network theory: applications to functional brain connectivity. *Netw Neurosci* **1**: 69–99.

Unpingco, J. (2019). *Python for Probability, Statistics, and Machine Learning*, 2, Switzerland: Springer International Publishing. https://www.springer.com/gp/book/9783030185442.

van der Maaten, L.J.P. (2013) *Barnes-Hut-SNE*, (pp. 1–11). Ithaca, New York, NY: arXiv. http://arxiv.org/abs/1301.3342.

Van Der Maaten, L. (2014) Accelerating t-SNE using tree-based algorithms. *J Mach Learn Res* **15**: 1–21.

Venter, C., Remington, K., Heidelberg, J., Halpern, A., Rusch, D., Eisen, J., *et al*. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**: 66–74.

Villaverde, A.F., Ross, J., Morán, F., and Banga, J.R. (2014) MIDER: network inference with mutual information distance and entropy reduction. *PLoS One* **9**: e96732.

Voigt, A., and Almaas, E. (2019) Assessment of weighted topological overlap (wTO) to improve fidelity of gene co-expression networks. *BMC Bioinform* **20**: 58.

Voorhies, A.A., Mark Ott, C., Mehta, S., Pierson, D.L., Crucian, B.E., Feiveson, A., Oubre, C.M., Torralba, M., Moncera, K., Zhang, Y., Zurek, E., and Lorenzi, H.A. (2019). Study of the impact of long-duration space missions at the International Space Station on the astronaut microbiome. *Scientific Reports* **9**. http://dx.doi.org/10.1038/s41598-019-46303-8.

Wadsworth, W.D., Argiento, R., Guindani, M., Galloway-Pena, J., Shelbourne, S.A., and Vannucci, M. (2017) An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. *BMC Bioinform* **18**: 94.

Wang, H., Li, M., Wang, J., and Pan, Y. (2011) *A New Method for Identifying Essential Proteins Based on Edge Clustering Coefficient* (pp. 87–98). Berlin, Heidelberg: Springer.

Washburne, A.D., Silverman, J.D., Leff, J.W., Bennett, D.J., Darcy, J.L., Mukherjee, S., *et al*. (2017) Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. *PeerJ* **5**: e2969.

Waskom, M., Botvinnik, O., Ostblom, J., Lukauskas, S., Hobson, P., M. Gelbart, *et al*. (2020) *seaborn*.GitHub. https://github.com/mwaskom/seaborn.

Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., *et al*. (2013) The cancer genome atlas Pan-cancer analysis project. *Nat Genet* **45**: 1113–1120.

Weiss, H., Hertzberg, V.S., Dupont, C., Espinoza, J.L., Levy, S., Nelson, K., and Norris, S. (2019). The airplane cabin microbiome. *Microbial Ecology* **77**: 87–95. http://dx.doi.org/10.1007/s00248-018-1191-3

Wetterstrand, K.A.(2020). *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)*, Maryland: National Human Genome Research Institute. https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data.

Winter, C., Kristiansen, G., Kersting, S., Roy, J., Aust, D., Knösel, T., *et al*. (2012) Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes. *PLoS Comput Biol* **8**: e1002511.

Yip, A.M., and Horvath, S. (2006). *The Generalized Topological Overlap Matrix For Detecting Modules in Gene Networks*, Las Vegas, NV: Proceedings of the 2006 International Conference on Bioinformatics & Computational Biology. https://www.semanticscholar.org/paper/The-Generalized-Topological-Overlap-Matrix-for-in-Yip-Horvath/6bff956be4ecdd433974821930b454a68fab1fac.

Yu, X., Zhang, J., Sun, S., Zhou, X., Zeng, T., and Chen, L. (2017) Individual-specific edge-network analysis for disease prediction. *Nucleic Acids Res* **45**: e170.

Zhang, L., Andersen, D., Roager, H.M., Bahl, M.I., Hansen, C.H.F., Danneskiold-Samsøe, N.B., *et al*. (2017) Effects of gliadin consumption on the intestinal microbiota

and metabolic homeostasis in mice fed a high-fat diet. *Sci Rep* **7**: 44613.

Zhang, B., and Horvath, S. (2005). A General Framework for Weighted Gene Co-Expression Network Analysis. *Statistical Applications in Genetics and Molecular Biology* **4**. http://dx.doi.org/10.2202/1544-6115.1128.

Zhou, W., Sailani, M.R., Contrepois, K., Zhou, Y., Ahadi, S., Leopold, S.R., *et al*. (2019) Longitudinal multi-omics of host–microbe dynamics in prediabetes. *Nature* **569**: 663–671.

Espinoza, J.L. (2020). *Compositional: Compositional data analysis in Python*, GitHub. https://github.com/jolespin/compositional.

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Predicting antimicrobial mechanism-of-action from transcriptomes: A generalizable explainable artificial intelligence approach

Josh L. Espinoza[1,2☯], Chris L. Dupont[1☯*], Aubrie O'Rourke[1], Sinem Beyhan[1], Pavel Morales[1], Amy Spoering[3], Kirsten J. Meyer[4], Agnes P. Chan[5], Yongwook Choi[5], William C. Nierman[5], Kim Lewis[4], Karen E. Nelson[1,2,5]

1 J. Craig Venter Institute, La Jolla, CA, United States of America, **2** Department of Applied Sciences, Durban University of Technology, Durban, South Africa, **3** NovoBiotic Pharmaceuticals, Cambridge, MA, United States of America, **4** Department of Biology, Northeastern University, Boston, MA, United States of America, **5** J. Craig Venter Institute, Rockville, MD, United States of America

☯ These authors contributed equally to this work.
* cdupont@jcvi.org

Citation: Espinoza JL, Dupont CL, O'Rourke A, Beyhan S, Morales P, Spoering A, et al. (2021) Predicting antimicrobial mechanism-of-action from transcriptomes: A generalizable explainable artificial intelligence approach. PLoS Comput Biol 17(3): e1008857. https://doi.org/10.1371/journal.pcbi.1008857

Editor: Avner Schlessinger, Icahn School of Medicine at Mount Sinai, UNITED STATES

Received: July 29, 2020

Accepted: March 8, 2021

Published: March 29, 2021

Data Availability Statement: Sequencing data for the E. coli transcriptomes challenged with various antibiotics were deposited using SRA identifiers SRR8909189 - SRR8909443 under BioProject PRJNA532938. Raw read counts and pairwise log2FC profiles are available in S7 and S8 Tables, respectively. Algorithm details are available in the S1 Methods. The soothsayer Python package, Clairvoyance command-line executables, and reproducible code in the form of Jupyter notebooks are available at https://github.com/jolespin/

Abstract

To better combat the expansion of antibiotic resistance in pathogens, new compounds, particularly those with novel mechanisms-of-action [MOA], represent a major research priority in biomedical science. However, rediscovery of known antibiotics demonstrates a need for approaches that accurately identify potential novelty with higher throughput and reduced labor. Here we describe an explainable artificial intelligence classification methodology that emphasizes prediction performance and human interpretability by using a Hierarchical Ensemble of Classifiers model optimized with a novel feature selection algorithm called *Clairvoyance*; collectively referred to as a CoHEC model. We evaluated our methods using whole transcriptome responses from *Escherichia coli* challenged with 41 known antibiotics and 9 crude extracts while depositing 122 transcriptomes unique to this study. Our CoHEC model can properly predict the primary MOA of previously unobserved compounds in both purified forms and crude extracts at an accuracy above 99%, while also correctly identifying darobactin, a newly discovered antibiotic, as having a novel MOA. In addition, we deploy our methods on a recent *E. coli* transcriptomics dataset from a different strain and a *Mycobacterium smegmatis* metabolomics timeseries dataset showcasing exceptionally high performance; improving upon the performance metrics of the original publications. We not only provide insight into the biological interpretation of our model but also that the concept of MOA is a non-discrete heuristic with diverse effects for different compounds within the same MOA, suggesting substantial antibiotic diversity awaiting discovery within existing MOA.

Author summary

As antimicrobial resistance is on the rise, the need for compounds with novel targets or mechanisms-of-action [MOA] are of the utmost importance from the standpoint of public health. A major bottleneck in drug discovery is the ability to rapidly screen candidate

soothsayer and open-sourced under the BSD-3 license.

Funding: This work was supported by the National Institute of Allergy and Infectious Diseases grant P01AI118687 to KL, ALS, and KEN. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: The following authors, ALS, and KL, declare competing financial interests as they are employees and consultants of NovoBiotic Pharmaceuticals.

compounds for precise MOA activity as current approaches are expensive, time consuming, and are difficult to implement in high-throughput. To alleviate this bottleneck in drug discovery, we developed a human interpretable artificial intelligence classification framework that can be used to build highly accurate and flexible predictive models. In this study, we investigated antimicrobial MOA through the transcriptional responses of *Escherichia coli* challenged with 41 known antibiotic compounds, 9 crude extracts, and a recently discovered (circa 2019) compound, darobactin, with novel MOA activity. We implemented a highly stringent Leave Compound Out Cross-Validation procedure to stress-test our predictive models by simulating the scenario of observing novel compounds. Furthermore, we developed a versatile feature selection algorithm, *Clairvoyance*, that we apply to our hierarchical ensemble of classifiers framework to build high performance explainable machine-learning models. Although the methods in this study were developed and stress-tested to predict the primary MOA from transcriptomic responses in *E. coli*, we designed these methods for general application to any classification problem and open-sourced the implementations in our *Soothsayer* Python package. We further demonstrate the versatility of these methods by deploying them on recent *Mycobacterium smegmatis* metabolomic and *E. coli* transcriptomics datasets to predict MOA with high accuracy.

Introduction

As antibiotic resistant pathogens have increasingly emerged [1,2], the discovery of new antimicrobials has lagged [3,4] despite previous efforts in screening hundreds of thousands of compounds [5]. Many of the screened compounds are either identical to known drugs, close analogs thereof, or have the same molecular targets [6]. Despite the wide variety of utilized antibiotics, many of these appear to collapse into 6 distinct mechanisms-of-action [MOA] based on pure enzyme inhibition assays. Progress within antibiotic discovery has been relatively slow [7] and the discovery of new antibiotics within existing MOA [8] constitutes a vanishingly low percentage of screened compounds [3,4]. This redundant discovery and diminishing returns of new chemical entities underpins declining industry efforts in screening for new antimicrobial drugs and a desire for disruptive new approaches.

One of the barriers to finding new chemical entities with novel biological targets is the problem of dereplication; the determination of a compound's primary MOA is time-consuming and often results in rediscovering a previously observed compound from a known MOA. The typical screening for antibiotics entails bacterial growth inhibition assays followed by macromolecular synthesis assay [9], with the former defining antibacterial activity and the latter determining the primary MOA [7]. Growth inhibition assays are easily automated and performed in high throughput [10]. An automated method to screen new antimicrobial compounds in high-throughput for both predicted MOA and similarity to known antibiotics as an intermediary step would obviate a major bottleneck in the path from drug discovery to clinical trials. Efforts to utilize more detailed whole cell bioreporter methods include large scale mutant library screening [11], whole cell imaging [12–14], proteome profiling [15,16], transcriptomics [12,17–20], and metabolomics [21]. Relative to the other approaches, transcriptome profiling benefits from capturing broad gene expression information relative to input labor. In previous MOA predictive modeling studies, accuracy estimated was occasionally absent [20], difficult to reproduce [19], or lacked robustness on held-out compounds [19]. More recent studies did validate their models but did not evaluate their models on unobserved

compounds [12,20,21]. The desired method should have high prediction accuracy validated on compounds not included in the training data and, therefore, unobserved by the model. An approach that abides by this stringency is Leave Compound Out Cross-Validation [LCOCV] where all instances of a compound are reserved for a testing set and the remaining compounds are used for model training; thus, demonstrating predictive performance on unobserved compounds. Even though a large set of transcriptomic data has been accrued in this field, the datasets have not been utilized effectively to build predictive MOA classification models, presenting a unique opportunity for exploiting recent advancements in artificial intelligence [AI] and machine learning.

As AI broadly mimics the cognitive abilities of the human mind, machine learning, a subset of AI, focuses on the ability of machines to receive input and adapt to information for a variety of tasks including predictive modeling and data mining for diagnostic genes. Machine-learning algorithms require large amounts of high-quality training data from intelligently designed experiments to effectively learn latent patterns that describe phenomena; in the case of this study, patterns within differential gene expression [DGE] profiles that can discriminate MOA. However, many high-performance models such as deep neural networks are difficult to interpret in a biological context where transparency in diagnostic decisions are paramount for reliable clinical applications. Explainable AI, often abbreviated XAI, is an effort to produce human interpretable models while maintaining a high level of learning performance [22]. Interpretability in the context of AI translates to a detailed understanding of a model's decision-making process. Although XAI cannot directly explain hitherto unknown biological phenomena, it can be used synergistically to guide research endeavors with domain expertise which, in turn, produce more realistic models resulting in a positive feedback-loop of information gain.

Given the proper training data, XAI can be leveraged to address two major questions of biotechnological and fundamental importance in antibiotic discovery. First, can XAI utilize whole transcriptome responses to predict the primary MOA of a compound or culture extract with high accuracy? In this scenario, antibacterial compounds or extracts that defy classification potentially represent new chemical entities with novel molecular targets or MOA; a major goal of biomedical science. Of seemingly lesser but potentially greater impact, does an examination of these responses within compounds of the same MOA reveal established MOA categories to be discrete entities or rather a spectrum of biological responses? In the latter case, compounds categorized within existing MOA but with unique transcriptional responses may represent new chemical entities that would have been discarded erroneously using traditional approaches.

Results

Antimicrobial mechanism of action training data compounds and producer-strain extracts

Our training dataset consists of 41 antibiotic majority FDA-approved compounds representing 6 MOA including inhibitors for protein-, DNA-, RNA-, cell-wall-, cell-membrane-, and fatty-acid-synthesis (Tables 1 and S1), which were chosen to maximize coverage of MOA and chemoinformatic space. The challenge experiments were conducted in *Escherichia coli* strain *W0153*, which has a permeable outer-membrane susceptible to large hydrophobic antibiotics [23], allowing us to investigate the effects of more antibiotic compounds at lower concentrations than wild-type strains, therefore, reducing the likelihood of off-target effects that could trigger secondary MOA activities. For each compound, at least triplicate challenges were conducted and transcriptomes were sequenced to analyze gene expression profiles.

Table 1. Training data for pure compounds and producer-strain extracts relative to MOA. Number of compounds, samples, and pairwise DGE profiles for pure compounds and producer-strain extracts relative to individual MOA.

	Pure Compounds			Producer-strain Extracts		
	Compounds	Samples	Pairwise DGE Profiles	Compounds	Samples	Pairwise DGE Profiles
MOA						
cell-membrane	2	11	33	0	0	0
cell-wall	12	61	178	4	18	54
dna-synthesis	10	52	171	2	7	21
fatty-acid-synthesis	3	12	36	0	0	0
protein-synthesis	9	42	126	2	6	18
rna-polymerase	4	20	58	2	6	18

https://doi.org/10.1371/journal.pcbi.1008857.t001

Historically, the majority of antibiotics have been discovered and isolated by fermenting soil bacteria. Hence for nine compounds, we also included crude extracts from organisms producing a specific antibiotic compound (called "producer-strain extracts" herein) to prepare our models for high-throughput discovery pipelines of microbial extracts, which would obviate the time-consuming chemical purification.

The specific machine learning problem addressed in this study is to robustly predict the MOA of a compound unobserved by the model using gene expression data generated from microbes treated with said compound. An added constraint of this overarching task is to ensure maximum model interpretability without sacrificing model performance and these objectives are evaluated by simulating predictive performance on novel compounds *in silico* (i.e., LCOCV). As machine learning algorithms benefit greatly from more high-quality training data, we used pairwise DGE profiles (instead of summary statistics) to maximize the number of observations while simultaneously accounting for bias between sampling and providing prediction error profiles. This simple procedure increased our available training data from 235 observations to 713 observations and, thus, providing more information that can be used for modeling (Table 1). With these 713 pairwise DGE profiles, we used 3065 protein-coding genes as features to increase opportunities for downstream interpretability and potential *post hoc* validation experiments.

Feature selection to optimize held-out compound classification performance

Machine learning models tend to overfit when the number of features vastly exceeded the number of observations; in this case, genes and biological samples, respectively. The training data dimensionality is not ideal for even simple binary classification models, let alone 6 imbalanced classes, thus, it was not surprising to find that most traditional classification models performed poorly (<90% LCOCV accuracy) (Table 2). Our solution to overcome this dimensionality obstacle was to develop the *Clairvoyance* feature selection algorithm as a means for curating gene sets that could robustly discriminate the primary MOA of DGE profiles. The objective function implemented in *Clairvoyance* maximizes the accuracy of custom (or stochastic) cross-validation pairs by iteratively enriching the subset of predictive features (e.g., genes). This iterative enrichment denoises the dataset with respect to a specific classification task resulting in a smaller feature set with reduced potential for model overfitting (see *S1 Methods*). In the case of this study, the *Clairvoyance* algorithm iteratively refines the gene sets to maximize the MOA classification accuracy of unobserved compounds provided as the test set in our custom LCOCV pairs to simulate the performance on novel compounds.

Table 2. Model performance using several supervised machine-learning algorithms. Various machine-learning algorithms were evaluated using the entire feature set (n = 3065 genes) and the *Clairvoyance*-optimized feature set ($GeneSet_{y1\text{-}y5}$, n = 399 genes) with the same LCOCV pairs. Performance metrics for each LCOCV set include accuracy, precision, recall, and F1 score. LCOCV refers to Leave Compound Out Cross Validation where we remove all instances of a compound from the data used to fit the model (training data) and evaluate performance on the held-out compound profiles (testing data) (see *Materials and Methods*).

	Clairvoyance feature selection [N = 399 Genes]				No feature selection [N = 3065 Genes]			
	Accuracy	F1 Score	Precision	Recall	Accuracy	F1 Score	Precision	Recall
Classifier								
CoHEC	0.999	0.983	0.983	0.982	0.749	0.693	0.715	0.682
Logistic Regression	0.880	0.829	0.856	0.817	0.793	0.732	0.763	0.723
Random Forest	0.792	0.719	0.768	0.708	0.742	0.659	0.703	0.645
K-Nearest Neighbors	0.714	0.568	0.617	0.546	0.636	0.506	0.561	0.481
Support Vector Machine	0.798	0.722	0.778	0.704	0.694	0.616	0.668	0.600
Naive Bayes (Gaussian)	0.698	0.582	0.623	0.561	0.429	0.302	0.389	0.274
AdaBoost	0.333	0.308	0.333	0.301	0.339	0.277	0.333	0.261
Neural Network	0.872	0.785	0.815	0.773	0.741	0.635	0.683	0.619

https://doi.org/10.1371/journal.pcbi.1008857.t002

We leveraged *Clairvoyance* feature selection with a multiclass version of a logistic regression model predicting MOA using a one-vs-rest architecture. Without feature selection, this model predicts the MOA from unobserved compounds with a LCOCV accuracy of 79.3% (Table 2). With feature selection designed for multiclass predictions, *Clairvoyance* was able to identify 98 genes ($GeneSet_{Multiclass}$) that could predict MOA from unobserved compounds with a LCOCV accuracy above 95% (Table 3). Although the performance of this model is high, we wanted to extend our methods to a hierarchical framework to better understand the decision-making process and maximize the amount of available information.

Hierarchical framework for multiclass classifications

With inspiration from the mechanisms of human cognition and the applications to automated facial recognition [24], we sought to decompose the complex task of multiclassification into a multilayered path of simple binary tasks [25]. We have developed a flexible framework for implementing Hierarchical Ensemble of Classifiers [HEC] models and their *Clairvoyance*-optimized counterpart [CoHEC]. Our basic HEC model approach implements a hierarchical ensemble of binary classifiers through a single graphical model with 3 degrees of flexibility for each sub-model decision node: (1) a custom feature set optimized for a simple binary classification task; (2) a unique classification algorithm with hyperparameters that most effectively discriminates the sub-model-specific decision paths; and (3) the relationship between sub-models can be data-driven or assigned *a priori*.

The graphical structure of our CoHEC model (Figs 1A and S2) is entirely data-driven to demonstrate the autonomous abilities of our XAI methodology by solely using emergent patterns within the training dataset in relation to the labeled classes. In other words, we do not predefine the graphical structure or gene sets using curated databases or domain knowledge (although, this functionality is supported) and instead allow the data to guide such parameter choices. Optimization of the gene feature set for each sub-model using *Clairvoyance* ($GeneSet_{yk}$ where k ranges from sub-models 1–5) boosted LCOCV accuracy substantially; between 10–23% in most cases and all cases resulting in left-out compound accuracies greater than 99% (Fig 1B and S2 Table). Several estimators were evaluated, optimized, and tuned for each sub-classification task but logistic regression models were the exemplar in all cases. While a few genes are shared between various pairs of sub-models, none of the 399 unique genes from $GeneSet_{y1\text{-}y5}$ used in the CoHEC model were universal to all sub-models reinforcing the notion

Table 3. Evaluating external datasets using CoHEC models. MOA prediction accuracy and performance when applying our methods to the data from Zoffmann et al. 2019 and Zampieri et al. 2018 and the methods from Hutter et al. 2004 on our dataset. In all cases, LCOCV was used for evaluating model performance for each individual observation (e.g. pairwise DGE profile), each cross-validation set (e.g. held out teixobactin), and using various majority voting schemes (see *Materials and Methods*). CPD is an abbreviation for compound. *Indicates protein-synthesis sub-MOA classification (30S/50S).

Dataset	Model	Organism	Feature Set Label	Feature Type	Features	MOA	CPD	Individual Pairwise Profiles Accuracy	LCOCV Test Set Accuracy	Majority Voting (Hard) Accuracy	Majority Voting (Soft) Accuracy	Data Source
This study (All MOA)	CoHEC	Escherichia coli (W01573)	GeneSet_y1-y5	Gene	399	6	41	0.9972	0.9986	1	1	https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA532938
This study (All MOA)	Clairvoyance-optimized multiclass logistic regression	Escherichia coli (W01573)	GeneSet_y1-y5	Gene	399	6	41	0.85714286	0.88017911	0.86440678	0.89830508	https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA532938
This study (30S/50S)	Clairvoyance-optimized binary logistic regression	Escherichia coli (W01573)	GeneSet_30S/50S	Gene	7	2*	9	0.9691358	0.96153846	1	1	https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA532938
This study (All MOA)	Clairvoyance-optimized multiclass logistic regression	Escherichia coli (W01573)	GeneSet_Multiclass	Gene	98	6	41	0.95936	0.967735	0.983051	1	https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA532938
This study (All MOA)	Support vector machine (Hutter et al. 2004 Methods)	Escherichia coli (W01573)	-	Gene	-	6	41	0.758	-	-	-	-
Zoffmann et al. 2019	CoHEC	Escherichia coli (BW25113)	GeneSet_Zoffmann	Gene	35	4	16	1	1	1	1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE110137
Zampieri et al. 2018 (*reference_t0*)	CoHEC	*Mycobacterium smegmatis*	MetaboliteSet_Zampieri-t0	Metabolite	492	18	62	0.949	0.949	0.977	0.991	https://www.ebi.ac.uk/biostudies/studies/S-BSST113
Zampieri et al. 2018 (*reference_solvent*)	CoHEC	*Mycobacterium smegmatis*	MetaboliteSet_Zampieri-solvent	Metabolite	494	18	62	0.882	0.882	0.954	0.963	https://www.ebi.ac.uk/biostudies/studies/S-BSST113

https://doi.org/10.1371/journal.pcbi.1008857.t003

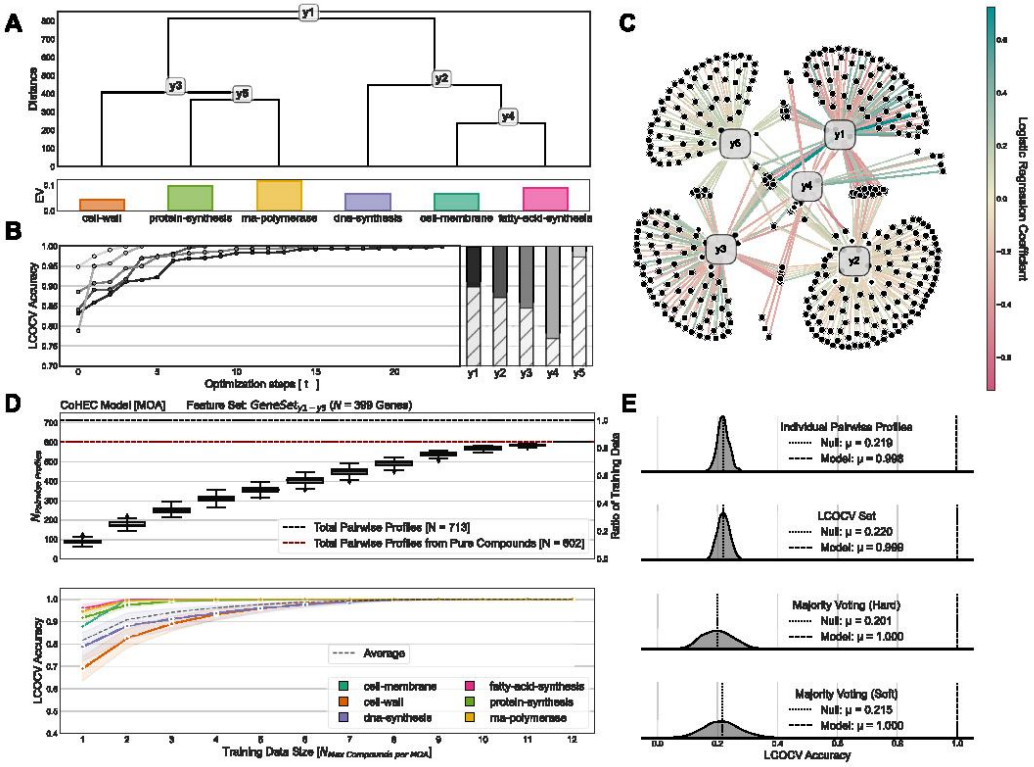


Fig 1. MOA classification performance and model benchmarking. A) The empirically determined structure of the CoHEC model calibrated to predict the MOA of an unobserved antibacterial compound based on the transcriptional change profiles of *E. coli*. The colored bar chart below the dendrogram shows the explained variance of the eigenprofile for each MOA. B) The influence of the *Clairvoyance* optimization algorithm for feature selection on model performance at each of the 5 sub-model decision points. Optimization step ($t = 0$) corresponds to using all available gene features, while each optimization step removes low information features during each consecutive iteration. The column chart shows the original baseline accuracy (lower) with all 3065 gene features and the effects of *Clairvoyance* optimized feature selection (upper). C) Network visualization of genes feature sets, determined by *Clairvoyance*, used by each sub-model decision point of the CoHEC model. The edge width represents the coefficient magnitude in each fitted Logistic Regression sub-model with the sign reflected by the color (positive = teal, negative = rose). D) Benchmarking of CoHEC model performance (N = 500 permutations without repetition) showing (upper) the number of compounds included during (lower) LCOCV evaluation relative to performance. Error bars represent standard error of mean. E) Kernel density of LCOCV accuracy for CoHEC null model (N = 500 permutations without repetition) and dashed horizontal lines representing actual CoHEC model performance.

https://doi.org/10.1371/journal.pcbi.1008857.g001

that each sub-model is task specific (Fig 1C and S2 and S3 Tables). Of these 399 genes in our CoHEC model, there were 87 of the 98 genes (88.8%) in $GeneSet_{Multiclass}$ overlapping (S6 Fig) and, thus, demonstrating the ability of *Clairvoyance* to identify emergent patterns within the data despite different model architectures. Interestingly, none of the MOA enzymatic targets were selected by *Clairvoyance* as discriminative features further endorsing our data-driven approach because the discriminating patterns were unknown *a priori*.

Our hierarchical framework provides a seamless avenue for introducing additional classification layers *post hoc* to a fitted model. For example, our CoHEC model was initially designed to classify 6 MOA categories but we've augmented the model with an additional layer to

predict sub-MOA for 30S/50S subunit protein-synthesis inhibitors to showcase this functionality. In our CoHEC model, protein-synthesis is discriminated from rna-polymerase inhibitors by sub-model *y5* using a subset of 92 genes ($GeneSet_{y5}$). To demonstrate how the information content in our hierarchical model is nested, we used *Clairvoyance* with the 92 genes from sub-model *y5* to identify an additional feature set that could robustly discriminate 30S from 50S protein-synthesis inhibitors (S1 Table). This approach predicted the target subunit of protein-synthesis inhibitors with a LCOCV accuracy greater than 96% using a subset of only 7 genes ($GeneSet_{30S/50S}$) (S5E and S5F Fig and S2 and S3 Tables) from sub-model *y5* and only overlapped with gene sets from sub-models in protein-synthesis decision path.

Model evaluation and benchmarking

The structure of our training data and our representation of differential expression allowed us to evaluate unobserved compound accuracy on 3 hierarchical abstractions. For our CoHEC model using $GeneSet_{y1\text{-}y5}$, we have the following evaluation: (1) the accuracy of individual pairwise DGE profiles using LCOCV (99.72%), (2) the mean accuracy for each LCOCV test set (99.86%), and (3) the majority voting consensus prediction for a compound from multiple individual predictions (100%) as shown in Tables 2,3,S4, and S7, and described in detail with *Materials and Methods*. Consensus predictions from our CoHEC model using individual predictions grouped by LCOCV test sets can be accomplished via soft majority voting with sub-model probabilities or hard majority voting with only terminal predictions; majority voting is a method that combines the results of multiple predictions into a single prediction. Regardless of the voting scheme, the CoHEC model achieves 100% accuracy for predicting the primary MOA from unobserved compounds despite the draconian method of leaving out all instances of a compound when fitting the model during LCOCV (S3 Fig and Tables 3 and S7). This methodology out-performs that of previous studies [14,19,21] despite our usage of a far more stringent accuracy validation method.

We compared our CoHEC model performance to similar models and methods to assess performance gains. As mentioned prior, we used *Clairvoyance* with a multiclass logistic regression and obtained accuracies greater than 96% using $GeneSet_{Multiclass}$ (Tables 2 and S7). To test whether our CoHEC model can outperform a standard multiclass model given the same input data, we evaluated another multiclass logistic regression but instead of using $GeneSet_{Multiclass}$ we used the 399 genes from $GeneSet_{y1\text{-}y5}$ which was designed for a hierarchical architecture. The multiclass version of our CoHEC model performed with a LCOCV accuracy between 85.7% and 89.8% depending on evaluation method (Tables 3 and S7). Although hierarchical feature selection that is designed for a multiclass model ($GeneSet_{Multiclass}$) or adapted from a CoHEC model ($GeneSet_{y1\text{-}y5}$) improves the classification performance when compared to a standard multiclass model without feature selection (LCOCV accuracy = 79.3%, Table 2), these methods cannot compete with the synergy of feature selection and hierarchical classifications implemented in our CoHEC model.

In addition to evaluating (Tables 3 and S7) and benchmarking (Fig 1D and 1E) our CoHEC model's ability to predict MOA, we also tested the performance of the following models: (1) multiclass logistic regression model using $GeneSet_{y1\text{-}y5}$ predicting MOA (S5A and S5B Fig); (2) *Clairvoyance*-optimized multiclass logistic regression model ($GeneSet_{Multiclass}$) predicting MOA (S5C and S5D Fig); and (3) the *Clairvoyance*-optimized binary logistic regression model predicting 30S/50S protein-synthesis inhibitors ($GeneSet_{30S/50S}$) (S5E and S5F Fig). The null LCOCV accuracy of our MOA predictive models had a similar range 20% - 24% (Figs 1E, S5B and S5D) which is only slightly above the expected null accuracy of 16.6% given perfect randomness.

By gradually increasing the number of compounds used for training, we were able to characterize the LCOCV accuracy distribution to evaluate how many compounds were needed to properly train the model (i.e., saturation) and if a model is overfitting. We define saturation in this context as a model's ability to robustly predict held-out compounds and stabilize even upon addition of more compounds into the training data. In particular, we observed a stark difference between the multiclass logistic regression model using $GeneSet_{y1\text{-}y5}$ and our CoHEC model using the same feature set, In particular, the multiclass representation had an initial LCOCV accuracy of 56.7% (± 2.93%) fitting the model with a single compound per MOA and does not ever saturate as each additional compound results in notable gains in performance with a maximum LCOCV accuracy of 99.1% using a maximum all 12 available compounds per MOA and all of the 713 pairwise DGE profiles (S5A Fig). In contrast, our CoHEC model using the same feature set, attained an initial LCOCV accuracy of 81.7% (± 2.72%) fitting the model with a single compound per MOA and surpasses the multiclass model's performance using a maximum of only 7 compounds per MOA upon saturation with an average of 448/713 pairwise DGE profiles (Fig 1D). Put simply, given the same amount of information, CoHEC models can learn predictive patterns faster and more robustly than the direct multiclass adaptation. With this, the CoHEC model surpasses the multiclass adaptation performance using 37% less data. Although our *Clairvoyance*-optimized multiclass logistic regression models fit using $GeneSet_{Multiclass}$ could predict MOA with high LCOCV accuracy (> 96%), we observed a lower benchmarking performance than our CoHEC with an initial LCOCV accuracy of 67.4% (± 2.78%) using a single compound per MOA and did not observe classification saturation until about 10 compounds per MOA. Our CoHEC model can outperform its multiclass counterparts and, therefore, derive more meaning given the same input data.

Interpreting trained models

Interpretability of trained models is paramount in XAI and CoHEC models provides substantial insight into the decision-making process. For instance, fitted HEC models produce an array of probabilities for each of the 5 sub-models (*y1-y5*) with built-in methods designed to calculate the probability for traversing each of the 10 decision paths and to visualize the predictions via decision graphs (Fig 2 and S4 Table). In this case, the probabilities represent binary decision paths from each of the 5 logistic regression sub-models (though other algorithms for sub-models are supported) and the standard error is calculated for profiles grouped by LCOCV test set; that is, all associated pairwise DGE profiles corresponding to a compound in a LCOCV test set. These 10 probabilities computed by the CoHEC model on LCOCV test sets are machine informative as unsupervised analysis of these probabilities clusters compounds by MOA with statistically greater homogeneity than the input data of pairwise DGE profiles (Figs 3B, 3D and S1); further shown when comparing silhouette score distributions (Fig 3C). The ability of our CoHEC model to compute probabilities that can confidently cluster a compound with its respective producer-strain extract in a completely unsupervised setting provides a powerful avenue to dereplicate known compounds in high throughput (Fig 3D and 3E).

The CoHEC model was field-tested by examining crude extracts from producer-strains for 9 compounds as we would implement in a practical antimicrobial discovery pipeline. Even when the respective pure-compound had not been observed by the model in the training set during our LCOCV procedure, the classifier accurately predicts these producer-strain extracts and this holds true for texiobactin as well; a recently discovered inhibitor of cell-wall biosynthesis [8] (Figs 2A, 2B and S3, and S4 Table). We observed an agreement of probabilities and standard errors for the prediction paths between a pure compound and the associated producer-strain extract suggesting the model is resilient to potential noise from metabolites

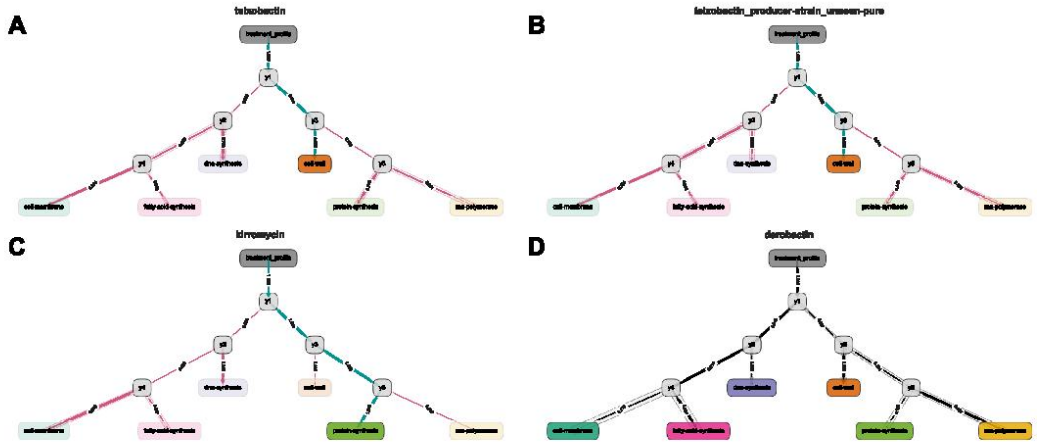


Fig 2. CoHEC model decision graphs for pure compounds, producer extracts, and darobactin representing MOA predictions. Prediction paths where each terminal colored node depicts a MOA, each internal gray node represents a sub-model decision point, and the edge-width corresponds to the probability according to the model for the respective path. Opaque halos around the edges represent SE with a large width corresponding to higher variance and vice versa. Rose and teal colored edges illustrate predictions traversing incorrect and correct paths, respectively, with black edges representing paths within a novel MOA paradigm. (A,B) Show teixobactin as a pure compound and the respective producer-strain while (C) depicts kirromycin and (D) represents darobactin. All of the prediction paths shown have no instance of the compound being previously observed by the model.

https://doi.org/10.1371/journal.pcbi.1008857.g002

present in the extracts; further supporting dereplication applications. This ability to predict primary MOA at the level of crude fermentation extract relieves the bottleneck of purification and isolation of active compounds in natural product antibiotic discovery, addressing our objectives of providing a high throughput method for primary MOA determination and compound dereplication. Our CoHEC model was also able to accurately predict kirromycin, a known protein-synthesis inhibitor via EF-Tu [26], even though it was not used during our training process (Fig 2C). Finally, a compound with a newly discovered MOA not present in the training data, darobactin [27], was examined. Simulating a novel MOA is difficult because the paradigms are entirely dependent on the input data, each having unique properties, but our examination of standard error profiles reveal a method for identifying novel MOA (Fig 2D). While the CoHEC probabilities for darobactin point to a fatty-acid-synthesis inhibitor, the standard error profile along the predicted path is the highest observed in the entire dataset, particularly at sub-model *y4* in discriminating between fatty-acid-synthesis and cell-membrane inhibitors (Fig 3E). The routing, albeit error prone, towards fatty-acid-synthesis and cell-membrane inhibition is biologically relevant as darobactin uniquely targets the β-barrel assembly machinery, the BAM complex, which is necessary for outer membrane protein biogenesis [28]. This one-off novel target prevents proper modeling of robust cutoffs in standard error for rejecting a prediction. However, this instance proves that a negative result contains immense value and can be leveraged for identifying new chemical entities with novel activity. When the CoHEC model fails to confidently classify an antibacterial compound, assuming proper data preprocessing, it likely has a novel MOA or target. While the CoHEC model has high accuracy at predicting primary MOA for known compounds, it also proves robust when identifying compounds within a MOA and compounds representing new MOA such as kirromycin and darobactin, respectively.

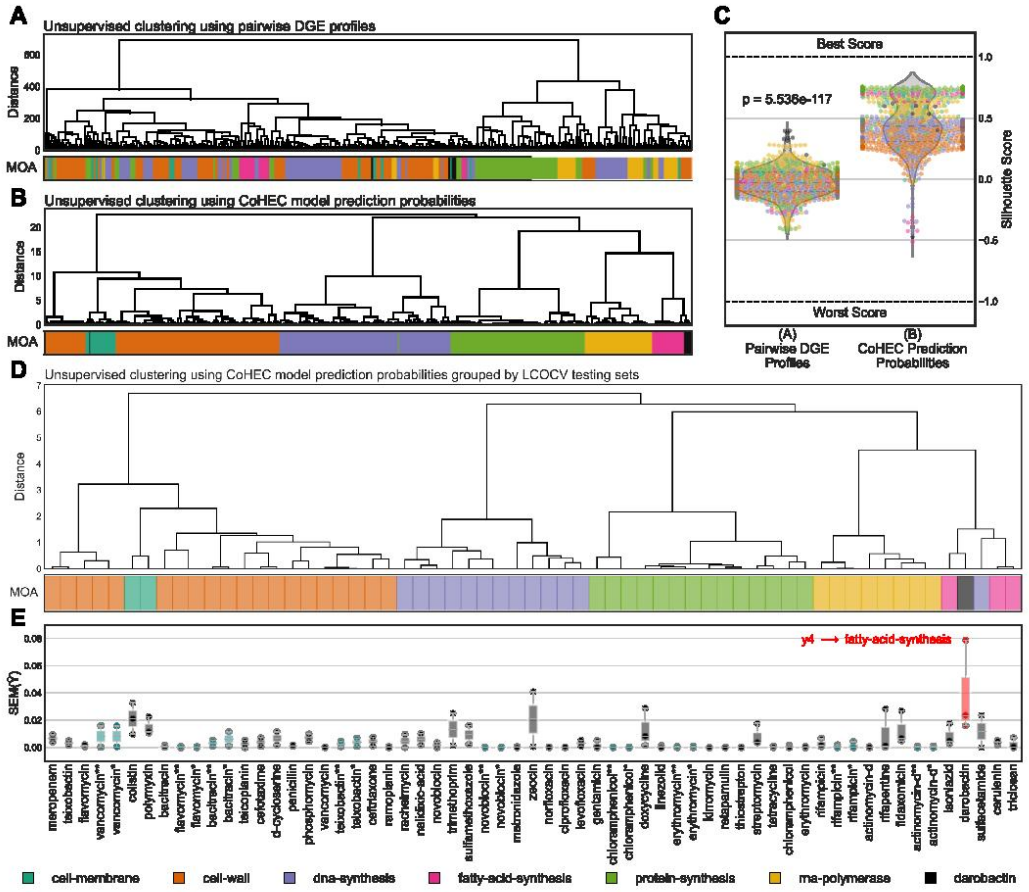


Fig 3. Unsupervised clustering performance and error profiles of transcriptomes and CoHEC model probability vectors. Unsupervised hierarchical clustering using (A) pairwise DGE profiles prior to feature selection (N = 3065 genes), (B) CoHEC model LCOCV test set prediction probabilities concatenated for all sub-models, and (D) CoHEC model prediction probabilities averaged by LCOCV test set. All hierarchical clustering uses Euclidean distance and ward linkage. C) Distributions of silhouette scores for (A) and (B) clustering results with Wilcoxon signed-rank test for statistical significance. D) Unsupervised hierarchical clustering and E) standard error profiles for each of the sub-model and the predicted path with red showing darobactin as a novel MOA and teal showing producer-extracts. Producer-strain extracts where the pure compound: (*) has been observed; and (**) has not been observed by the HEC model in the training data. The box plots extend from the Q1 to Q3 quartile values of the data, with a line at the median (Q2), and whiskers at 1.5 * IQR.

https://doi.org/10.1371/journal.pcbi.1008857.g003

Interpreting models based on gene expression data is difficult as this approach often captures downstream effects. Regardless, the decision graphs and sub-model gene coefficients are biologically relevant when evaluated via *Gene Set Enrichment Analysis* [GSEA] [29]. For instance, coefficient-ranked genes from sub-model *y2* (DNA-synthesis vs. *y4*) are enriched in both DNA and membrane-related GO terms (GO:0009432, GO:0006281, GO:0009102, GO:0006974, GO:0090305, GO:0009314, GO:0004518, GO:0006310, GO:0003677) while *y4* (cell-membrane vs. fatty-acid-synthesis) is enriched in membrane-related (GO:0006810, GO:0005886, GO:0016020, GO:0016021) and transport (GO:0006810) GO terms as shown in

S5 Table. We also observed several nucleotide-binding related (GO:0003677, GO:0000166), transcription regulation (GO:0006355), protein-binding (GO:0005515, GO:0042802), and metal-binding (GO:0000287, GO:0046872, GO:0008270, GO:0051539) GO terms enriched in sub-model *y5* in the classification between protein-synthesis and rna-polymerase inhibitors.

Extending methodology to datasets of other microbial strains and feature modalities

To evaluate our methods relative to other published antibiotic discovery work, we used a collection of transcriptomics and metabolomics datasets classifying MOA utilizing different microbial strains and feature modalities than used in this study. Hutter et al. 2004 generated a database of *Bacillus subtilis* transcriptional responses to treatments of 37 well-characterized antibacterial compounds from different MOA which were used to build a support vector machine model to predict MOA of antibacterial compounds. The training data from Hutter et al. 2004 [19] was not published in any public database. However, the support vector machine modeling approach and data transformations were well-documented so we used their methodology on our dataset to compare method performance. The methods from Hutter et al. 2004 applied to our dataset resulted in 54.5% (no normalization), 60.6% (TMM normalization), and 75.8% ($\log_2$ transformation) LCOCV accuracies (Tables 3 and S6), which is substantially lower than our CoHEC model. However, the Hutter et al. 2004 methodology used an unconventional approach that concatenates samples with respect to the feature axis, thus, increasing feature dimensionality and lowering the numbers of observations available for training. We used a more standard approach (i.e., stacking replicates on the observation axis instead of feature axis) in implementing support vector machines to evaluate the performance using modern methodology but this only increased the LCOCV accuracy by 3% (Table 2).

Next we tested our methods on external datasets by re-analyzing the transcriptomic data from Zoffmann et al. 2019 [12] and metabolomic data from Zampieri et al. 2018 [21]. Zoffmann et al. used a combination of transcriptomics and cell imaging data to predict 7 MOA in a different *E. coli* strain (see *Materials and Methods*). Zoffmann et al. did not publish the cell imaging data used to construct predictive models but we were able to download the counts from *NCBI Gene Expression Omnibus* (Accession: GSE110137) consisting of *E. coli* BW25113 challenged with 16 compounds. However, because of this inability to access the same data the Zoffmann et al. models cannot be directly compared to the results in our study. With the available public data, we computed pairwise DGE profiles, built CoHEC models, and optimized each sub-model using *Clairvoyance* with the same protocol and commands used to construct the CoHEC model in this study. The CoHEC model for the Zoffmann et al. 2019 transcriptomic data resulted in 100% LCOCV accuracy using only 35 gene features (S6 Table). However, Zoffmann et al. 2018 implemented a random forest classifier which we also implemented as shown in Table 2; though, CoHEC models out-performed this method and other standard classifiers.

The Zampieri et al. 2018 study had the most complete data that was publicly available, accessible through *EMBL-EBI BioStudies* (Accession: S-BSST113)[21]. This study used an iterative hypergeometric test to model metabolite responses of *Mycobacterium smegmatis* exposed to 62 compounds representing 18 MOAs. The Zampieri et al. metabolomic data had both a temporal aspect and contained a reference solvent for each timepoint. We constructed two CoHEC model paradigms: (*reference_t0*) treatment at $\boldsymbol{t_n}$ vs. treatment at $\boldsymbol{t_0}$; and (*reference_solvent*) treatment at $\boldsymbol{t_n}$ vs. solvent at $\boldsymbol{t_n}$ as both are biologically informative. We adapted our LCOCV strategy to incorporate treatment concentration for MOA that contained only a single representative. With this dataset, our CoHEC models achieved a LCOCV between 94.9% - 99.1% LCOCV accuracy with 492 metabolite features when using $t = 0$ as a reference and

88.2% - 96.3% with 494 metabolite features when using solvent as a reference (Tables 3, S6 and S7). Zampieri et al. 2018 [21] reports their performance using area under the curve, which is undefined for LCOCV, thus we were not able to directly compare model performance.

Discussion

The CoHEC models present a purely data-driven XAI approach that can predict the primary MOA from unobserved compounds with high performance. This data-driven AI maximizes the available information content by asking simple questions about specific genes in a particular order to effectively evade statistical artifacts that are inherent in biological datasets where features greatly exceed the number of observations. We demonstrate the resourcefulness of our CoHEC methodology by comparing multiclass models either using the same input data ($GeneSet_{y1\text{-}y5}$) or with gene sets designed for multiclass models ($GeneSet_{Multiclass}$) and evaluating the number of compounds needed per MOA to stabilize prediction performance. The CoHEC model can exceed the performance of a multiclass model using the same base algorithm (e.g., logistic regression in this study) with only a fraction of the training data when using the same input features. This is desired in the field of bioinformatics where sample collection is a limiting factor and interpretability is key.

Furthermore, our hierarchical classification scheme is intuitive in that we can visualize the flux of weighted decisions through the graph for both individual and grouped observations (Figs 2 and S3 and S3 and S4 Tables). Most importantly, our approach does not sacrifice performance for interpretability because the CoHEC model can be unpacked to reveal feature weights that directly translate the AI decision process into human comprehensible terminology.

As membrane/transport GO terms were expected to be enriched in gene sets that classify MOA targets related to cellular structure and nucleotide/protein binding related terms were expected for gene product synthesis, we were not expecting a multitude of metal ion related GO terms in the classification of protein-synthesis and rna-polymerase inhibitors. However, this agrees with previous studies that have focused on metal-responsive ECF sigma factors, several of which are activated by iron depletion or by an excess of other metals such as zinc [30]; thus, overlapping with the GO terms enriched in our GSEA analysis (S5 Table). Bacterial ECF sigma factors are directly involved in the transcription process by recognizing promoter sequences, together with the core RNA polymerase enzyme, and initiate the transcription of the genes they regulate [31]. Although our models can be fully understood from a mathematical perspective, biological interpretation is limited to previous empirical studies and the extent of domain knowledge available. However, our methods are expected to provide a powerful resource in guiding empirical validation experiments to demystify complex biological processes.

While some multiclass classification problems do not require the architecture of hierarchical methods (e.g., S1A Fig), many more likely do given that negative data-mining results are rarely published. Our methods allow each decision to be evaluated and optimized with flexibility in classification algorithms, custom cross-validation-based objective functions, feature selection optimization, and hyperparameter tuning for each sub-model (S2 Table). In addition, the estimators of each sub-model could be further incorporated into ensemble methods such as tree-based gradient boosting for non-linear discrimination (e.g., XGBoost [32], CatBoost [33]) or AdaBoost [34] with logistic regression to further boost performance. Ultimately, the implementations developed here are widely adaptable to a variety of research goals where mining descriptive features for discriminating groups or complex classifications are desired. For instance, *Clairvoyance* was developed and validated on primary antibiotic MOA predictions

but nascent versions of this algorithm were implemented to identify genes and pathways associated with cyanobacteria-moss symbiotic events [35], some of which have been experimentally validated *post hoc* using gene knockout experiments, demonstrating broad usage. We further demonstrate versatility by applying our methods to predict MOA from *E. coli BW25113* transcriptomics and *M. smegmatis* metabolomics timeseries (Tables 3,S6 and S7). In these demonstrations, we reveal that our AI determined a mix of logistic regression and non-linear tree-based classifiers to be optimal for predicting MOA. Furthermore, we expand the methods to investigate metabolomic profiles both in relation to solvent (*reference_solvent*) and to a baseline timepoint prior to antibiotic treatment (*reference_t_0*) showcasing how one could investigate MOA from different biological contexts. Our methods outperformed these studies regardless of strain, species, modality, or large number of MOA categories while using a robust LCOCV accuracy metric.

Many empirical MOA classification experiments are based on macromolecular synthesis assays with limited targets and response variables. An open scientific question prior to this study was whether compounds with the same MOA determined by enzyme assays elicit the same whole transcriptome response. The lack of consensus statistically significant differentially expressed genes within a primary MOA (S1C Fig), the disparity of the global transcriptome in response to compounds both within and between MOA (Fig 3A), and the intermixing in multivariate analyses (S1B Fig) all suggest that concept of a MOA is a non-discrete fuzzy categorization. Although this presents challenges for classification algorithms, it also illuminates that there is unexplored functional space of new chemical entities within existing MOA that needs to be surveyed as compounds within a MOA can have very different biological effects. The 3 MOA with the lowest number of compounds in this study, cell-membrane (n = 2), fatty-acid-synthesis (n = 3), and RNA-polymerase (n = 4) are underrepresented in our training data because there is a very limited number of FDA-approved compounds (Tables 1 and S1) for these MOA. With the proper experimental design, our methodology could identify novel targets within these underrepresented MOA and expand the map of each MOA landscape and, in doing so, our understanding of antimicrobial resistance as a whole. In addition, a natural extension to our discovery methods would be to build secondary *post hoc* model as a successive layer to the main CoHEC model able to determine if and how a successfully classified compound is functionally divergent from previously observed compounds. A companion study examining sub-MOA diagnostic features may address the research required to execute such an addition [36]. In future work, we plan to empirically validate our model predictions and expand the underrepresented MOA classes to fortify the AI's understanding of MOA-specific patterns.

As we have demonstrated, the data-driven methods used here are designed to be transferrable to other organisms as our primary goal was to rapidly screen a broad range of compounds with any antimicrobial activity. The specific methods developed here need no specialized equipment beyond access to a sequencing core, which is near universal, but we also demonstrate usage in other modalities such metabolomics. This is a benefit over previous methods that required extensive numbers of genetically modified reporter strains [11], mass spectrometers [16,21], or high-end microscopes [13]. Ultimately, this method is easily utilized by other researchers and the algorithms have been designed to flexibly accommodate model updates by automating feature selection, determining hierarchical structure, and parameter tuning with parallel-computing scalability for use on personal laptops to high-performance compute servers. Progressive science is built on open-sourcing knowledge, which is why we used an inexpensive publicly available strain, a detailed experimental design, and hosted the algorithms with tutorials demonstrating usage in an open-sourced programming language. These are components that facilitate an organically collaborative community resource accelerating antimicrobial discovery in both biology-centric and data-driven paradigms. Given our conclusion

that there are likely unexplored spaces within existing MOA, such an effort should yield new chemical entities with novel activity and provide a unique perspective in data-mining for other researchers.

Materials and methods

Selecting antibiotic compounds

An initial set of antibiotics to be tested was chosen to represent the breadth of FDA-approved antibiotics across MOA classes and then certain MOA classes were supplemented with non-approved compounds with known antibiotic MOA to improve diversity. Subsequently, this compound set was dereplicated according to structural diversity using an ordination based on the molecular descriptors of the compounds [36].

Crude extract production of antimicrobial-producing microbial strains

To further test the predictive model capabilities, strains producing known antibiotics were fermented and the whole broth was processed to produce crude extracts as described previously in a parallel study to this research [36] (S1 Table).

Strains were inoculated from a frozen glycerol stock onto SMSR4 agar plates (0.125 g casein, 0.1 g potato starch, 1 g casamino acids, 100ml R4 fermentation medium, 20 g bacto-agar in 1 L water). Morphology was confirmed under a 10X magnification using a Zeiss Stemi 2000 microscope and inoculated into 20ml of Modsb (15 g glucose, 10 g malt extract, 10 g soluble starch, 2.5 g yeast extract, 5 g casamino acids, and 0.2 g $CaCl_2$-$2H_2O$ per 1 L deionized H_2O, pH 7.0) in a 250ml flask, shaken at 150 rpm at 28˚C for 2–5 days. Upon robust growth, the biomass was diluted 1:20 into 500ml of production medium R4 (10 g glucose, 1 g yeast extract, 0.1 g casamino acids, 3 g proline, 10 g $MgCl_2$-$6H_2O$, 4 g $CaCl_2$-$2H_2O$, 0.2 g K_2SO_4, 5.6 g TES free acid (2-[[1,3-dihydroxy-2-(hydroxymethyl)propan-2-yl]amino]ethanesulfonic acid) per 1 L deionized H_2O, pH 7) for all strains except X4251. X4251 was diluted 1:20 into 500ml of production medium BPM (20 g glucose, 10 g organic soy flour (Bob's Red Mill), 10 g pharmamedia (Traders Protein), 1 g $(NH_4)_2SO_4$, 10 g $CaCO_3$, 20 g glycerol per 1 L deionized H2O). Activity was monitored by bioassay and the active cultures were harvested between 4 and 7 days of growth in the production medium at 150 rpm at 28˚ C. Crude extracts were generated by extracting the whole broth culture with an equal volume of water saturated n-butanol for 3 hours at room temperature and sonicated in a water bath for 20 mins prior to clarifying the butanol/aqueous layers with centrifugation. The n-butanol layers were removed into clean tubes and dried in a Savant Speedvac Concentrator heated to 45oC under vacuum. The dried substances were reconstituted and concentrated in 100% DMSO at 10X the original volume. Crude extracts were divided into 500ul aliquots, tested for MIC against W0153, and kept frozen until used for exposures to produce transcriptomes. The production of known compounds was confirmed with mass spectrometry, HPLC retention time, and/or spectrum of activity against resistant and sensitive test strains. Crude extracts were shipped on dry ice from Novo-Biotic Pharmaceuticals to JCVI overnight.

Antibiotic challenge experiments and sequencing

Escherichia coli strain W0153 (parent strain AB1157; asmB1 tolC::kan modifications) was acquired from the Yale culture collection (http://cgsc2.biology.yale.edu/Strain.php?ID=4509). This modified AB1157 strain of *E. coli* has the *asmB1* allele, which reduces LPS synthesis, and the gene for *tolC* has been replaced by a Kanamycin resistance cassette. For the antibiotic challenge, 3 mls of *E. coli* strain W0153 at an $OD_{600}$of 0.5, representing mid-log phase, were

exposed to each antibiotic in biological triplicate at 1xMIC for 30 minutes. After 30 minutes of exposure, 100 μls of the cells were removed for OD_{600} values and CFU/ml counts (S4 Fig). This served as a checkpoint to observe that the 1xMIC antibiotic treated sample is showing an OD_{600} value and CFU/ml counts less than that than of the untreated control $t = 30$ minute solvent control but greater than that of the $t = 0$ sample, to ensure proper growth and to rule out an over treatment of the cells for an incorrect MIC. In parallel, the remainder of the cells were immediately pelleted at 4˚C by centrifugation for 10 minutes at 2000 rpm in 1ml aliquots. The supernatant was removed and the pellets were immediately frozen in liquid nitrogen then stored for the RNA extraction processing at a later date. Total RNA was extracted by automation using the NucleoMag RNA extraction kit (*Macherey-Nagel*, *GmbH*) on the EpMotion Robotic liquid handler. For the resulting total RNA, RIN values were obtained to check for RNA quality using the 2200 TapeStation (*Agilent Genomics*, *Inc*.). Acceptable values to proceed to ribosomal subtraction were above a RIN of 5. Ribosomal RNA (rRNA) was subtracted from the total RNA to yield only messenger RNA for library construction using a bacterial rRNA depletion kit (*New England Biolabs*, *Inc)* at half reactions with a total RNA input maximum of 400 ng. The rRNA depleted product was quality controlled using an Agilent Bioanalyzer with the Agilent Pico chip for RNA detection to check for less than 0.5% of rRNA remaining. Then, 2.5 μl of the rRNA depleted samples, amounting to approximately 2–5 ng, is used as the input material to construct each cDNA library for RNA sequencing using the NEBNext Ultra Directional RNA Library prep kit (*Illumina*, *Inc*.) at half reactions. The resulting libraries were analyzed using Agilent High Sensitivity DNA chips to ensure library quality. Libraries were quantified and normalized by qPCR and then sequenced using the NextSeq 500 High Output Kit at 150 cycles producing approximately 9 million, 75 base-pair, paired-end reads for each library.

Sequence processing, mapping, and normalization

Reads were quality trimmed and mapped to *E. coli* K-12 substr. MG1655 (Genbank: U00096.2, EcoCyc: v21.1) using *clc* (http://resources.qiagenbioinformatics.com) to produce a gene counts matrix (S8 Table). To maximize the number of observations and capture all of the variance in our dataset we used pairwise DGE profiles of the Trimmed Mean of M-values [TMM] normalized counts after filtering out a subset of genomic features. We removed low-quality samples that had fewer than 4000 detected genes or less than 100,000 reads mapping to non-ribosomal genes. The following genes were removed from the remaining samples: (1) genes other than rRNA whose abundance were sensitive to ribosomal depletion methods [G26 (D-galactose 1-dehydrogenase), G0-8867 (GcvB small regulatory RNA), EG30069 (RnpB RNA), G0-9281 (glutamate-pyruvate aminotransferase), and EG30100 (tmRNA)]; (2) rRNA genes; (3) non-protein-coding genes; (4) genes differentially expressed between comparisons of media and antibiotic carrier controls; and (5) genes differentially expressed in response to the producer-strain metabolic background (i.e. pure compound vs. producer extract). Our method of pairwise DGE is calculated by the following: (1) remove genes described above; (2) TMM normalization using *edgeR* [37]; and (3) for each compound in a sequencing run we calculate the $\log_2(\text{compound}_r)$—$\log_2(\text{control}_{r'})$ for all compound replicates $\boldsymbol{r}$ and respective control replicates $\boldsymbol{r'}$ using a pseudocount of 1 (S9 Table). Our dataset consists of 9 sequencing runs, each with several antibiotics representing different MOA, and we only include relationships within a sequencing run to minimize batch effects and reduce variance introduced from non-biological processes. Statistically significant differentially expressed genes were computed using *edgeR*'s exactTest with $|\log_2 FC| \geq 2$ and FDR < 0.001 to minimize the influence of off-target effects (S1C Fig).

Hierarchical ensemble of classifiers modeling

The graphical structure of our HEC model is entirely data-driven to exploit natural patterns within the data. However, it is possible to use a predefined structure but, due to the limitations in our understanding of latent variables in biological classification tasks, we implemented an unsupervised method to allow the data to dictate the hierarchy. Our methods for implementing this unsupervised hierarchy alludes to the concept of an *eigengene* which, essentially, is the first principal component of a dataset using a subset of features [38]. In this context, we transpose the operation by generating *m*-dimensional *eigenprofiles* representing each MOA class from our pairwise DGE feature matrix [***X***] (m = the number of genes). We then use classical agglomerative methods with Euclidean distance and ward linkage to cluster these profiles revealing the relationships between MOAs as a natural hierarchical structure entirely dependent on the differential expression profiles. The implementation for this pipeline can be found within the soothsayer.hierarchical.Topology object.

Once the structure is determined, the framework resembling a decision-tree is used to construct a directed *NetworkX* [39] directed graph where each internal or terminal node in the graph represents a sub-model or classification category (e.g. MOA), respectively. The *paths* for each classification target in the directed graph and the target matrix [***Y***] can be obtained from the soothsayer.hierarchy.Topology object using the get_paths and get_target_matrix methods, respectively. The resulting *paths* contain a collection of ordered nodes when traversing the graph towards the desired target classification from the input node. ***Y*** contains the binary classifications for each sub-model in the graph and is used with ***X*** to train all of the sub-models simultaneously using the fit method. The model object is implemented in soothsayer.classification.HierarchicalClassifier and mimics the API of *scikit-learn* (arXiv:1309.0238).

Each sub-model node serves as a vessel for storing custom fields including feature sets (e.g., gene subsets from feature selection), feature matrices (e.g., gene expression or pairwise DGE data), and *scikit-learn* compatible classification models among other custom data fields. The edge weights between nodes in the graph contain probabilities from the parent sub-models and these can be examined quantitatively or visualized for a qualitative assessment of a single prediction or group of predictions (e.g., replicates) with standard error of the mean error bar support. The sub-model nodes contain a unique classification model equipped with custom model hyperparameters and gene subsets designed to optimize a specific classification task in the overarching model. The sub-model hyperparameters and gene sets are shown in S2 Table.

This architecture allows maximum flexibility for decomposing complex predictions into a sequence of simple predictions with any set of features or any *scikit-learn* compatible classification model. The implementations for preprocessing data, determining hierarchical structures, building and evaluating hierarchical classification models, and analyzing diverse datasets can be found within our *Soothsayer* Python package. Additional supervised-classifier algorithms, as shown in Table 2, were implemented and evaluated using *scikit-learn* with random_state = 0 when applicable.

An example of the CoHEC prediction process for the transcriptomic response to teixobactin (Fig 1C): (1) evaluate using the 102 genes in sub-model *y1* with 99% probability diffusing towards sub-model *y3*; and (2) sub-model *y3* uses a subset of 101 genes, of which only 10 genes overlap with sub-model *y1*, routing the transcriptome profile to the cell-wall MOA with 99.5% probability as a terminal classification.

Simulating novel compounds by evaluating model on unobserved compounds via Leave Compound Out Cross Validation

In the context of drug discovery, our LCOCV training and testing splits simulated the following scenarios: (1) if there is only a pure-compound then we leave out all profiles for the

compound in the test set to simulate an unobserved compound; (2) if there are both producer-extracts and pure compounds we (2a) leave out all profiles related to the compound and test only on the pure compound (again, simulating an unobserved compound); (2b) leave out all profiles related to a compound and test only the producer-extract (simulating an unobserved compound derived from extract); and (2c) leave out only the producer-extract from the test set (simulating a known compound derived from extract). With this scheme, we end up with 59 unique LCOCV training and testing splits (Fig 3D and 3E and S4 Table).

Feature selection (Clairvoyance)

Clairvoyance is a novel feature selection algorithm designed to enrich a dataset for features that maximize an accuracy-based objective function. In the case of this study, the feature selection is applied to pairwise DGE profiles to identify gene sets that optimize classification accuracy for the specific binary classification task associated with each sub-model. The methods in *Clairvoyance* extend on concepts inspired by Zakharov and Dupont 2011 [40] and Warshan et al. 2017 [35] by adding pseudo-random sampling preserving class proportions, iterative processes, subsetting feature weights from classifiers with an accuracy threshold, and the use of both decision tree- and logistic regression-based ensembles for versatile performance. *Clairvoyance* implements parallel computations that are scalable and can be configured for running quickly on local machines for notable performance gains or exhaustively on compute clusters for even greater boosts in performance. The *Clairvoyance* algorithm is available in our *Soothsayer* Python package implemented as 1) a low-level object for prototyping in interactive consoles as soothsayer.feature_extraction.Clairvoyant and (2) a stand-alone executable with very few dependencies.

The objective function of *Clairvoyance* maximizes the cross-validation accuracy which can accept custom cross-validation training/testing pairs; in this case, leaving all instances of a compound out for a testing set (i.e., LCOCV). In the context of drug discovery, our objective function was to maximize LCOCV accuracy of held-out compounds to simulate the performance on novel compounds unobserved by the model.

The basic strategy of *Clairvoyance* is as follows: (1) iterate through a grid of hyperparameter configurations and for each iteration $\boldsymbol{k}$ construct classifier $\boldsymbol{clf_k}$; (2) shuffle the training data without replacement into equally sized observation subsets $\boldsymbol{A}$ and $\boldsymbol{B}$ while maintaining class proportions to produce $(\boldsymbol{X_A}, \boldsymbol{y_A})$ and $(\boldsymbol{X_B}, \boldsymbol{y_B})$ training/testing pairs, respectively; (3) train $\boldsymbol{clf_k}$ on $(\boldsymbol{X_A}, \boldsymbol{y_A})$ and predict on $(\boldsymbol{X_B}, \boldsymbol{y_B})$; (4) train $\boldsymbol{clf_k}$ on $(\boldsymbol{X_B}, \boldsymbol{y_B})$ and predict on $(\boldsymbol{X_A}, \boldsymbol{y_A})$; (5) store the weights (i.e. $|coefficients|$ for logistic regression or *feature importances* for tree-based models) of each feature (e.g. gene), the accuracy of the held out subset, and the hyperparameters of $\boldsymbol{clf_k}$ for the fitted models from steps (3 and 4); and repeat steps (1–5) $\boldsymbol{n\_iter}$ times for each hyperparameter configuration $\boldsymbol{k}$. More specific algorithm details can be found in the S1 Methods and in the open-sourced code.

The hyperparameter grid for logistic regression includes the following: (C) inverse of regularization strength; and (penalty) the penalization type as either *L1* or *L2* regularization. For decision tree classifiers, we include the following hyperparameters: (criterion) the function to measure the quality of a split with *gini* for Gini impurity or *entropy* for the information gain; (max_features) the number of features to consider when looking for the best split; and (min_samples_leaf) the minimum number of samples required to be at a leaf node.

The weights for each fit are collected in an array and reduced into a single weight vector with features listed in descending ordered by their predictive capacity. Each feature is iteratively added and the classifier is cross-validated using either custom training/testing sets or randomly generated stratified *K*-fold splits serving as the objective function maximized by

Clairvoyance. An ***early_stopping*** parameter is used to stop the algorithm from adding features if there have not been advancements in the accuracy for a user-specified number of iterations (100 in this study) to increase computational efficiency. Summary statistics are generated for the cross-validation performance of each feature subset and plots are generated showing the transitions between feature subsets with respect to cross-validation classification accuracy.

The basic form of the algorithm can be augmented by running the cross-validation methods on models with accuracy levels above a particular threshold, using either logistic regression and/or tree-based methods, and iteratively feeding enriched subsets into the algorithm for exhaustive data-mining to maximize the performance of the feature selection. These methods can be configured for single run use or in a pipeline that runs all configurations and produces a synopsis of all executions sorted by the highest performing run. A major component of the algorithm's flexibility is the incorporation of both logistic regression and decision trees for the objective function maximization as some discrimination tasks are better described by linear relationships of log-odds while other by non-linear criteria. The resulting feature subsets can be further explored using ensemble methods such as random forests and boosting ensembles with Bayesian or randomized hyperparameter tuning. We used the gene sets derived from *Clairvoyance* to build individual sub-models in our HEC model.

Clairvoyance identified several combinations of gene sets and hyperparameters of equally high accuracy using the 41-compound set listed in S1 Table (not including kirromycin or darobactin). To determine which gene sets and hyperparameter configurations would be used for each sub-model, we sorted each configuration by the following criteria and in this order: (1) ↓ LCOCV accuracy for 41 training compounds, (2) ↑ μ(standard error) of predicted path for 41 training compounds, and (3) ↑ number of genes; ↑ (lower is better) and ↓ (higher is better) refer to ascending and descending order, respectively.

Evaluation methodologies and benchmarking

The hierarchical nature of our dataset allowed us to evaluate our methodology in multiple ways. Our data is arranged in the following hierarchy: pairwise DGE profile → transcriptome → compound → MOA as shown in Table 1. With this hierarchy, we are able to use LCOCV training and testing splits to evaluate our data. As we perform LCOCV, we stack the predictions for all of the test sets that were held out into an array and once we complete cross-validation we can compute the overall LCOCV accuracy we refer to as "Individual Pairwise Profiles Accuracy". When we compute the accuracy of each LCOCV test set, we can also compute average accuracy of all the sets which we refer to as "LCOCV Test Set Accuracy".

Our experimental design includes several replicates for each compound treatment. These replicates in combination with the pairwise DGE profiles result in several observations to predict for a given compound and can be grouped via majority-voting methods using either soft voting (averaged probabilities) or hard voting (only considering terminal predictions). In the case of this study, hard voting would translate to predicting each profile separately and then using the most common prediction while soft voting sums the sub-model probabilities for all of the replicates and averaging. Soft majority voting can be calculated by averaging the probability matrix $\hat{\mathbf{Y}}$ (Y_hat where rows are testing pairwise DGE profiles and columns are sub-model probabilities), ensuring the sub-model probabilities sum to 1.0, and traversing the path of highest probability. In Python, this operation is achieved by Y_hat.mean(axis = 0*)* where Y_hat is the output of the predict_proba method built into the HierarchicalClassifier object. This aggregated probability profile is used as input for the predict_from_probas method which yields the most probable path from a directed walk across the aggregated probabilities. Hard majority voting can be computed simply by computing the prediction with the most

occurrences via y_hat.value_counts().idxmax() where y_hat is a collection of terminal predictions. If the most common predicted MOA is not unique then the prediction is rendered inconclusive via hard voting. Both majority voting methods have proven to be equally robust for predicting left-out compound. Our dataset allows us to use these majority voting approaches since we have discrete groupings (compounds) in each of our broader classification categories (MOAs). This grouping method only works for categories with discrete subcategories and, thus, does not apply to all classification tasks.

To provide deeper insight into model fitting characteristics, we benchmarked our models using two separate approaches. The first of which was by fitting the model with a variable number of random compounds from each MOA and evaluating model performance on the entire LCOCV set (Figs 1D, S5A, S5C and S5E). The second approach was by shuffling the MOA labels and fitting these null models on the shuffled dataset (Figs 1E, S5B, S5D and S5F). For both approaches, we repeated this process for 500 iterations to obtain a distribution of values instead of single point estimates. We used the same 500 random seeds in our sampling for each of our 3 MOA models, which use the same pairwise DGE profiles albeit different gene sets as either $GeneSet_{y1\text{-}y5}$ or $GeneSet_{Multiclass}$, making our benchmarking results comparable between methodologies.

Gene set enrichment analysis

Sub-model specific gene sets were evaluated using the logistic regression coefficients from each fitted sub-model against the *Gene Ontology* database (http://geneontology.org/docs/download-ontology/ | go.obo [format 1.4, releases/2019-05-09]) using *GSEA's Prerank* module with 1000 permutations [29]. The gene sets were extracted from *EcoCyc* v21.1 flat files (data/gene_association.ecocyc). Significance is determined by FDR < 0.25 as suggested by the *GSEA* documentation.

External studies and datasets

We evaluated our methods using 3 external studies including Hutter et al. 2004 methods (no data available), Zoffmann et al. 2019 (transcriptomics), and Zampieri et al. 2019 (metabolomics). Each dataset contained their own caveats for analysis. Hutter et al. 2004 did not publish data but described modeling methodology. In this case, we were able to reproduce model methodology but not use the same data therefore we could not directly compare results. The Zoffmann et al. 2019 study did not publish cell imaging data used for modeling but did publish an auxiliary transcriptomics dataset that we were able to leverage for modeling. However, the Zoffmann et al. 2019 transcriptomics dataset had several MOA with only a single compound making the use of robust evaluation such as LCOCV impossible. To adapt their dataset to our stringency, we used slightly broader MOA categories by adjusting as follows: {"cell wall synthesis inhibitor / lipoprotein", "cell wall synthesis inhibitor"} → "cell-wall" and {"DNA replication inhibitor", "DNA damage", "Folic Acid synthesis inhibitor"} → "dna-synthesis" so we could have more than one representative per MOA. Zampieri et al. 2018 was the most comprehensive and accessible dataset. We were able to obtain metabolomic profiles but since we used LCOCV accuracy, we could not directly compare to their AUC scores as AUC is undefined for LCOCV. To properly integrate the zscore-normalized metabolic data with our methods, we used pairwise differences instead of pairwise DGE profiles as these were the most comparable.

Supporting information

S1 Fig. Unsupervised clustering and marker gene signatures. Model dataset (TCGA-PAN-CANCER, http://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq) and (B)

our training data from this study. A) Shows the discrimination of 5 different cancer types based on gene expression patterns for 801 samples with 20531 gene features. In this scenario, "out-of-the box" standard multivariate statistical analyses are sufficient to differentiate the cells types with high confidence. (B) Shows the results for the same multivariate analysis using the training data from this study which includes the transcriptional response to 41 compounds in 6 categories from 713 observations and 3065 gene features represented using Principle Component Analysis ordination. Multivariate statistical analyses cannot discriminate these samples by their MOA. (C) Differentially expressed genes [DEGs] shared by different compounds within a MOA. The low proportions (y-axis) shows that different compounds within a MOA have a different transcriptional response in terms of DEGs and that there is not a clear diagnostic profile for each MOA.
(EPS)

S2 Fig. XAI pipeline for determining hierarchical structure, selecting gene feature sets, and building MOA classifier. The pipeline begins with a basic next generation sequencing procedure for generating the training data for our method which includes treatment of organism with a compound of interest, transcriptome sequencing, read mapping to reference, filtering data, and generating pairwise DGE profiles (refer to *Materials and Methods*). The remainder of the pipeline is domain agnostic and is broadly applicable. The training data consists of a feature matrix **X** (e.g. the pairwise DGE profiles) and the target vector **y** (e.g. MOA classification). The training data is fed into *Soothsayer*'s *Topology* object which determines the hierarchical structure of the HEC model. Each of the sub-models at internal nodes within the tree-like structure undergo a feature selection procedure via *Soothsayer*'s *Clairvoyance* algorithm. This procedure determines a gene subset that optimizes the accuracy for the performance of the sub-model. Next, the: (1) sub-model estimators; (2) sub-model specific gene feature sets; and hierarchical structure are fed into *Soothsayer's Hierarchical Classifier* object to build the CoHEC model. This CoHEC model is a MOA classifier and can be validated by removing all instances of a compound from the training data, training the model on this subset, and then testing the model's MOA prediction accuracy with the left-out profile subset.
(EPS)

S3 Fig. MOA- and compound-specific model LCOCV prediction accuracy. MOA-specific prediction accuracy for unobserved compounds with a heatmap (left) showing the accuracy for each compound at each sub-model in the CoHEC model. Bar chart (right) showing the mean accuracy for each compound for the terminal prediction of the CoHEC model colored by MOA. Producer-strain extracts where the pure compound: (*) has been observed; and (**) has not been observed by the CoHEC model in the training data. Error bars reflect standard error of mean unless specifically noted otherwise.
(EPS)

S4 Fig. Survival rate for compounds. Percent survival for various compounds compared to solvent controls measured at $t = 30$ minutes via optical density of sample at wavelength of 600 nm. Error bars are standard deviations taken from 3 biological replicates with 3 technical replicates each. Crude extract indicated by [CE] suffix.
(EPS)

S5 Fig. Model performance and benchmarking for multiclass MOA and 30S/50S protein-synthesis classifiers. Benchmarking of *Clairvoyance*-optimized (A) multiclass logistic regression MOA and (B) binary 30S/50S protein-synthesis sub-MOA model performance (N = 500 permutations without repetition) showing (upper) the number of compounds included during (lower) LCOCV evaluation relative to performance. Error bars represent standard error of

mean. Kernel density of LCOCV accuracy for (B) multiclass MOA and (D) binary 30S/50S null models (N = 500 permutations without repetition) and dashed horizontal lines representing actual model performance.
(EPS)

S6 Fig. Gene set overlaps between MOA predictive model feature selection. Upset plots showcasing gene set overlap of CoHEC sub-models, multiclass MOA, and binary 30S/50S feature selection.
(EPS)

S1 Table. Antibiotic compound list with MICs and MOA categories. Each of the 41 antibiotics used in the training data with minimum inhibitory concentrations and MOA categorization.
(XLSX)

S2 Table. Sub-model parameters, features, and performance. Classifier parameters for *scikit-learn* estimators for each sub-model along with gene sets and performance metrics before and after optimization.
(XLSX)

S3 Table. Sub-model-specific gene sets with functional annotations and weights. Gene sets used CoHEC model derived from *Clairvoyance* feature selection with fitted logistic regression coefficients and *EcoCyc* annotations. Additionally includes multiclass MOA model and 30S/50S protein-synthesis inhibitor model coefficients.
(XLSX)

S4 Table. CoHEC model prediction probabilities for cross-validation and test sets. Prediction probability paths from cross-validation and test set combinations with respect to each sub-model. Cross-validation training and testing pairs for each sub-model included.
(XLSX)

S5 Table. Gene set enrichment analysis of fitted sub-model coefficients. Results from *GSEA's Prerank* module using coefficients as ranked weights with gene sets from the Gene Ontology database.
(XLSX)

S6 Table. Sub-model parameters, features, and performance for external datasets. Classifier parameters for *scikit-learn* estimators for each sub-model along with gene sets and performance metrics before and after optimization. These parameters pertain to Zoffmann et al. 2019 and Zampieri et al. 2018.
(XLSX)

S7 Table. Classification metrics. Classification metrics such as accuracy, f1 score, precision, and recall for different models and evaluation methods.
(XLSX)

S8 Table. Gene expression counts. Unnormalized gene expression counts.
(TSV)

S9 Table. Model training and testing dataset. Pairwise $\log_2$FC differential gene expression profiles.
(TSV)

S1 Methods. *Clairvoyance* algorithm. Detailed description of *Clairvoyance* algorithm including parameters and operations.
(DOCX)

Acknowledgments

We would like to acknowledge Suren Singh of *Durban University of Technology*, South Africa for his mentorship and support during this work.

Author Contributions

Conceptualization: Chris L. Dupont, Aubrie O'Rourke, Amy Spoering, William C. Nierman, Kim Lewis, Karen E. Nelson.

Data curation: Josh L. Espinoza.

Formal analysis: Josh L. Espinoza.

Funding acquisition: Amy Spoering, William C. Nierman, Kim Lewis, Karen E. Nelson.

Investigation: Josh L. Espinoza, Pavel Morales, Agnes P. Chan, Yongwook Choi.

Methodology: Josh L. Espinoza, Chris L. Dupont, Aubrie O'Rourke.

Project administration: Amy Spoering, William C. Nierman, Kim Lewis, Karen E. Nelson.

Resources: Chris L. Dupont, Aubrie O'Rourke, Sinem Beyhan, Amy Spoering, Kim Lewis, Karen E. Nelson.

Software: Josh L. Espinoza.

Supervision: Chris L. Dupont, Aubrie O'Rourke, Sinem Beyhan, Amy Spoering, Kim Lewis, Karen E. Nelson.

Validation: Josh L. Espinoza, Pavel Morales, Amy Spoering.

Visualization: Josh L. Espinoza.

Writing – original draft: Josh L. Espinoza, Chris L. Dupont, Aubrie O'Rourke.

Writing – review & editing: Josh L. Espinoza, Chris L. Dupont, Aubrie O'Rourke, Sinem Beyhan, Kirsten J. Meyer, William C. Nierman, Kim Lewis, Karen E. Nelson.

References

1. Davies J, Davies D. Origins and Evolution of Antibiotic Resistance. Microbiology and Molecular Biology Reviews. 2010; 74(3):417–33. https://doi.org/10.1128/MMBR.00016-10 PMID: 20805405
2. Laxminarayan R, Duse A, Wattal C, Zaidi AKM, Wertheim HFL, Sumpradit N, et al. Antibiotic resistance—the need for global solutions. The Lancet Infectious Diseases. 2013; 13(12):1057–98. https://doi.org/10.1016/S1473-3099(13)70318-9 PMID: 24252483
3. Baltz RH. Marcel Faber Roundtable: is our antibiotic pipeline unproductive because of starvation, constipation or lack of inspiration? Journal of industrial microbiology & biotechnology. 2006; 33(7):507–13. Epub 2006/01/19. https://doi.org/10.1007/s10295-005-0077-9 PMID: 16418869.
4. Group PCTRSW. A scientific roadmap for antibiotic discovery. 2016.
5. Payne DJ, Gwynn MN, Holmes DJ, Pompliano DL. Drugs for bad bugs: confronting the challenges of antibacterial discovery. Nat Rev Drug Discov. 2007; 6(1):29–40. https://doi.org/10.1038/nrd2201 PMID: 17159923.
6. Lewis K. The Science of Antibiotic Discovery. Cell. 2020; 181(1):29–45. https://doi.org/10.1016/j.cell.2020.02.056 PMID: 32197064

7. Silver LL. Challenges of antibacterial discovery. Clin Microbiol Rev. 2011; 24(1):71–109. https://doi.org/10.1128/CMR.00030-10 PMID: 21233508; PubMed Central PMCID: PMC3021209.
8. Ling LL, Schneider T, Peoples AJ, Spoering AL, Engels I, Conlon BP, et al. A new antibiotic kills pathogens without detectable resistance. Nature. 2015; 517:455. https://doi.org/10.1038/nature14098 https://www.nature.com/articles/nature14098#supplementary-information. PMID: 25561178
9. Cotsonas King A, Wu L. Macromolecular synthesis and membrane perturbation assays for mechanisms of action studies of antimicrobial agents. Current protocols in pharmacology. 2009;Chapter 13:Unit 13A.7. Epub 2009/12/01. https://doi.org/10.1002/0471141755.ph13a07s47 PMID: 22294390.
10. Hutter B, John GT. Evaluation of OxoPlate for real-time assessment of antibacterial activities. Curr Microbiol. 2004; 48(1):57–61. https://doi.org/10.1007/s00284-003-4095-4 PMID: 15018104.
11. Xu HH, Trawick JD, Haselbeck RJ, Forsyth RA, Yamamoto RT, Archer R, et al. Staphylococcus aureus TargetArray: comprehensive differential essential gene expression as a mechanistic tool to profile antibacterials. Antimicrob Agents Chemother. 2010; 54(9):3659–70. https://doi.org/10.1128/AAC.00308-10 PMID: 20547796; PubMed Central PMCID: PMC2934999.
12. Zoffmann S, Vercruysse M, Benmansour F, Maunz A, Wolf L, Blum Marti R, et al. Scientific Reports. 2019; 9(1):5013. https://doi.org/10.1038/s41598-019-39387-9 PMID: 30899034.
13. Nonejuie P, Burkart M, Pogliano K, Pogliano J. Bacterial cytological profiling rapidly identifies the cellular pathways targeted by antibacterial molecules. Proc Natl Acad Sci U S A. 2013; 110(40):16169–74. https://doi.org/10.1073/pnas.1311066110 PMID: 24046367; PubMed Central PMCID: PMC3791758.
14. Peach KC, Bray WM, Winslow D, Linington PF, Linington RG. Mechanism of action-based classification of antibiotics using high-content bacterial image analysis. Mol Biosyst. 2013; 9(7):1837–48. https://doi.org/10.1039/c3mb70027e PMID: 23609915; PubMed Central PMCID: PMC3674180.
15. Brotz-Oesterhelt H, Bandow JE, Labischinski H. Bacterial proteomics and its role in antibacterial drug discovery. Mass Spectrom Rev. 2005; 24(4):549–65. https://doi.org/10.1002/mas.20030 PMID: 15389844.
16. Mateus A, Bobonis J, Kurzawa N, Stein F, Helm D, Hevler J, et al. Thermal proteome profiling in bacteria: probing protein state in vivo. Mol Syst Biol. 2018; 14(7):e8242. https://doi.org/10.15252/msb.20188242 PMID: 29980614; PubMed Central PMCID: PMC6056769.
17. Shaw KJ, Miller N, Liu X, Lerner D, Wan J, Bittner A, et al. Comparison of the changes in global gene expression of Escherichia coli induced by four bactericidal agents. Journal of molecular microbiology and biotechnology. 2003; 5(2):105–22. Epub 2003/05/09. https://doi.org/10.1159/000069981 PMID: 12736533.
18. Wilson M, DeRisi J, Kristensen HH, Imboden P, Rane S, Brown PO, et al. Exploring drug-induced alterations in gene expression in Mycobacterium tuberculosis by microarray hybridization. Proc Natl Acad Sci U S A. 1999; 96(22):12833–8. https://doi.org/10.1073/pnas.96.22.12833 PMID: 10536008; PubMed Central PMCID: PMC23119.
19. Hutter B, Schaab C, Albrecht S, Borgmann M, Brunner NA, Freiberg C, et al. Prediction of mechanisms of action of antibacterial compounds by gene expression profiling. Antimicrob Agents Chemother. 2004; 48(8):2838–44. https://doi.org/10.1128/AAC.48.8.2838-2844.2004 PMID: 15273089; PubMed Central PMCID: PMC478524.
20. Jones MB, Nierman WC, Shan Y, Frank BC, Spoering A, Ling L, et al. Reducing the Bottleneck in Discovery of Novel Antibiotics. Microb Ecol. 2017; 73(3):658–67. https://doi.org/10.1007/s00248-016-0889-3 PMID: 27896376.
21. Zampieri M, Szappanos B, Buchieri MV, Trauner A, Piazza I, Picotti P, et al. High-throughput metabolomic analysis predicts mode of action of uncharacterized antimicrobial compounds. Science Translational Medicine. 2018; 10(429):eaal3973. https://doi.org/10.1126/scitranslmed.aal3973 PMID: 29467300
22. Gunning D. Explainable artificial intelligence (xai). Defense Advanced Research Projects Agency (DARPA), nd Web. 2017; 2(2).
23. Opperman TJ, Ling LL, Schumacher T, Puyang X, Moir DT. Novel permeable strains of E coli with improved properties for screening and mechanism of action studies. Abstracts of the Interscience Conference on Antimicrobial Agents & Chemotherapy. 2003; 43:255–. PMID: 035409856.
24. Su Y, Shan S, Chen X, Gao W. Hierarchical Ensemble of Global and Local Classifiers for Face Recognition. IEEE Transactions on Image Processing. 2009; 18(8):1885–96. https://doi.org/10.1109/TIP.2009.2021737 PMID: 19556198
25. Efstathiou J, Rajkovic V. Multi-attribute decision-making using a fuzzy heuristic approach. International Journal of Man-Machine Studies. 1980; 12(2):141–56.

26. Wolf H, Chinali G, Parmeggiani A. Kirromycin, an inhibitor of protein biosynthesis that acts on elongation factor Tu. Proc Natl Acad Sci U S A. 1974; 71(12):4910–4. https://doi.org/10.1073/pnas.71.12.4910 PMID: 4373734; PubMed Central PMCID: PMC434009.

27. Imai Y, Meyer KJ, Iinishi A, Favre-Godal Q, Green R, Manuse S, et al. A new antibiotic selectively kills Gram-negative pathogens. Nature. in review.

28. Han L, Zheng J, Wang Y, Yang X, Liu Y, Sun C, et al. Structure of the BAM complex and its implications for biogenesis of outer-membrane proteins. Nature Structural &Amp; Molecular Biology. 2016; 23:192. https://doi.org/10.1038/nsmb.3181 https://www.nature.com/articles/nsmb.3181#supplementary-information. PMID: 26900875

29. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005; 102(43):15545–50. https://doi.org/10.1073/pnas.0506580102 PMID: 16199517

30. Moraleda-Muñoz A, Marcos-Torres FJ, Pérez J, Muñoz-Dorado J. Metal-responsive RNA polymerase extracytoplasmic function (ECF) sigma factors. Mol Microbiol. 2019; 112(2):385–98. Epub 2019/06/26. https://doi.org/10.1111/mmi.14328 PMID: 31187912.

31. Helmann JD. The extracytoplasmic function (ECF) sigma factors. Adv Microb Physiol. 2002; 46:47–110. Epub 2002/06/21. https://doi.org/10.1016/s0065-2911(02)46002-x PMID: 12073657.

32. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; San Francisco, California, USA. 2939785: ACM; 2016. p. 785–94.

33. Dorogush AV, Ershov V, Gulin A. CatBoost: gradient boosting with categorical features support. arXiv. 2018.

34. Freund Y, Schapire RE. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Sciences. 1997; 55(1):119–39. https://doi.org/10.1006/jcss.1997.1504

35. Warshan D, Espinoza JL, Stuart RK, Richter RA, Kim SY, Shapiro N, et al. Feathermoss and epiphytic Nostoc cooperate differently: expanding the spectrum of plant-cyanobacteria symbiosis. ISME J. 2017; 11(12):2821–33. https://doi.org/10.1038/ismej.2017.134 PMID: 28800136; PubMed Central PMCID: PMC5702739.

36. O'Rourke A, Beyhan S, Choi Y, Morales P, Chan AP, Espinoza JL, et al. Mechanism-of-Action Classification of Antibiotics by Global Transcriptome Profiling. Antimicrobial Agents and Chemotherapy. 2020; 64(3):e01207–19. https://doi.org/10.1128/AAC.01207-19 PMID: 31907190

37. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26(1):139–40. Epub 2009/11/11. https://doi.org/10.1093/bioinformatics/btp616 PMID: 19910308.

38. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008; 9:559. https://doi.org/10.1186/1471-2105-9-559 PMID: 19114008; PubMed Central PMCID: PMC2631488.

39. Hagberg AA, Schult DA, Swart PJ. Exploring network structure, dynamics, and function using NetworkX. In: Varquaux G, Vaught T, Millman J, editors. Proceedings of the 7th Python in Science Conference (SyPy2008); Pasadena, CA USA2008. p. 11–5.

40. Zakharov R, Dupont P, editors. Ensemble Logistic Regression for Feature Selection2011; Berlin, Heidelberg: Springer Berlin Heidelberg.

EBioMedicine 73 (2021) 103644

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.elsevier.com/locate/ebiom

Interactions between fecal gut microbiome, enteric pathogens, and energy regulating hormones among acutely malnourished rural Gambian children

Helen M. Nabwera[a,†], Josh L. Espinoza[b,c,$], Archibald Worwui[d], Modupeh Betts[e], Catherine Okoi[d], Abdul K. Sesay[d], Rowan Bancroft[d], Schadrac C. Agbla[f], Sheikh Jarju[d], Richard S. Bradbury[g], Mariama Colley[d], Amadou T. Jallow[d], Jie Liu[h], Eric R Houpt[h], Andrew M. Prentice[d], Martin Antonio[d,i], Robin M Bernstein[j], Christopher L. Dupont[b,*], Brenda A. Kwambana-Adams[e,*]

[a] Department of Clinical Sciences, Liverpool School of Tropical Medicine, Liverpool, L3 5QA, UK
[b] J. Craig Venture Institute, 4120 Capricorn Ln, La Jolla, CA 92037, USA
[c] Applied Sciences, Durban University of Technology, Durban, South Africa
[d] Medical Research Council Unit The Gambia at London School of Hygiene and Tropical Medicine, Fajara, Banjul, PO Box 273, The Gambia
[e] NIHR Global Health Research Unit on Mucosal Pathogens, Division of Infection and Immunity, University College London, London, United Kingdom
[f] Department of Health Data Science, University of Liverpool, Liverpool, UK
[g] School of Health and Life Sciences, Federation University
[h] Division of Infectious Diseases and International Health, Department of Medicine, University of Virginia, Charlottesville, Virginia, United States of America
[i] Department of Infection Biology, Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, UK
[j] Growth and Development Lab, Department of Anthropology, University of Colorado, Boulder, CO, United States of America

ARTICLE INFO

Article History:
Received 26 July 2021
Revised 6 October 2021
Accepted 7 October 2021
Available online xxx

Keywords:
Malnutrition
Gut microbiome
Enteric pathogens
West Africa
Escherichia-Shigella
Network analysis
Feature selection
Community detection

ABSTRACT

Background: The specific roles that gut microbiota, known pathogens, and host energy-regulating hormones play in the pathogenesis of non-edematous severe acute malnutrition (marasmus SAM) and moderate acute malnutrition (MAM) during outpatient nutritional rehabilitation are yet to be explored.
Methods: We applied an ensemble of sample-specific (intra- and inter-modality) association networks to gain deeper insights into the pathogenesis of acute malnutrition and its severity among children under 5 years of age in rural Gambia, where marasmus SAM is most prevalent.
Findings: Children with marasmus SAM have distinct microbiome characteristics and biologically-relevant multimodal biomarkers not observed among children with moderate acute malnutrition. Marasmus SAM was characterized by lower microbial richness and biomass, significant enrichments in *Enterobacteriaceae*, altered interactions between specific *Enterobacteriaceae* and key energy regulating hormones and their receptors.
Interpretation: Our findings suggest that marasmus SAM is characterized by the collapse of a complex system with nested interactions and key associations between the gut microbiome, enteric pathogens, and energy regulating hormones. Further exploration of these systems will help inform innovative preventive and therapeutic interventions.
Funding: The work was supported by the UK Medical *Research Council (MRC; MC-A760-5QX00) and the UK Department for International Development (DFID) under the MRC/DFID* Concordat agreement; Bill and Melinda Gates Foundation (OPP 1066932) and the National Institute of Medical Research (NIMR), UK. This network analysis was supported by NIH U54GH009824 [CLD] and NSF OCE-1558453 [CLD].

1. Introduction

Protein-energy malnutrition [PEM] is defined as a range of pathological conditions arising from inadequate calories and/or protein intake [1]. PEM in early childhood (< 5 years) is endemic in low-and

* Corresponding: Dr Brenda Anna Kwambana-Adams, Christopher L. Dupont
E-mail addresses: cdupont@jcvi.org (C.L. Dupont), brenda.kwambana@ucl.ac.uk (B.A. Kwambana-Adams).
† Equal contribution

https://doi.org/10.1016/j.ebiom.2021.103644

Research in Context

Evidence before this study

There is increasing evidence that the pathogenesis of acute malnutrition may be multifactorial and linked to microbial dysbiosis, the overgrowth of specific enteric pathogens in the gut and hormonal imbalance. However, most of this evidence has been generated from children with edematous severe acute malnutrition [SAM] in East Africa and Southeast Asia. However, despite malnutrition being prevalent in West Africa, little is known about the pathogenesis of severe and moderate acute malnutrition during childhood in this region. Unlike East Africa and Southeast Asian, non-edematous nutrition dominates in this region, making it a unique case study.

Added value of this study

Here we provide novel insights into the role of microbial dysbiosis, enteric pathogens, and host energy-regulating hormones in the pathogenesis of both severe and moderate non-edematous acute malnutrition among West African Children during outpatient nutritional rehabilitation. This study also demonstrates key differences in the microbiome structure and enteric infections between children with severe and moderate non-edematous acute malnutrition.

Implications of all the available evidence

As with edematous severe acute malnutrition, non-edematous SAM is also characterized by the collapse of a complex system including the gut microbiome, enteric pathogens, and energy regulating hormones, but in with different important features. Future interventions should target these systems to improve the management and outcomes for children with acute malnutrition.

middle-income countries and is a contributory factor in up to 50% of under 5 mortality [2]. Acute malnutrition or wasting is a manifestation of PEM and is defined by the World Health Organization as weight-for-height Z scores [WHZ] below the median of the WHO Growth Reference Standards by at least 2 standard deviations [3]. It has varying degrees of severity from moderate acute malnutrition [MAM] to severe acute malnutrition [SAM] which can present as: 1) kwashiorkor or edematous SAM; 2) marasmus or non-edematous SAM; and 3) marasmic-kwashiorkor [4]. The 2018 Gambia *Multiple Indicator Cluster Survey* estimated an acute malnutrition prevalence of 7.4% among children between 6-23 months of age [5], while the Gambia *National Micronutrient Survey* estimated an acute malnutrition prevalence of 5.8% for children under 59 months of age [6].

SAM has multifactorial causes [7], which may contribute to the poor performance of nutritional interventions in combating this public health challenge [8]. In addition to inadequate infant feeding practices and household food insecurity that are invariably critical factors, acute malnutrition may also develop, progress, or persist as a result of enteric pathogens [9] and gut mucosal barrier dysfunction [10,11] associated with unhygienic living conditions from early childhood. In addition, the pathology of acute malnutrition may vary across different populations giving rise to similar nutritional phenotypes. Research investigating how the gut microbiome plays a role in acute malnutrition have been mostly limited to kwashiorkor and based in countries in East and Southern Africa, South-East Asia, and Central America [12-15] with substantial evidence that gut microbiome structure varies across regions [16,17]. While a previous metagenomic study has investigated acute malnutrition in the West African region [18], the focus was on the relationship between human milk oligosaccharide composition and microbiome composition.

The gut microbiome is complex and is only partially determined by the host genome. Several factors are involved in shaping the gut microbiome such as environmental modulation and diet [19,20]. For example, diets in low-and middle-income countries are often rich in complex plant polysaccharides compared to the energy-dense animal-derived foods and processed carbohydrates in high income countries [7]. Differences in dietary carbohydrate composition select for bacteria able to metabolize the available nutrients and such differences have been observed when comparing the fecal microbiota between healthy children in sub-Saharan Africa and those in Europe [21,22]. Depending on diet, gastrointestinal microbes produce a plethora of metabolites that modulate the host's fitness, phenotype, and health [23]. Given the variation in gut microbiota between individuals and the interplay between microbial and host metabolites in relation to nutrition, it might be possible for nutritional interventions aimed at acutely malnourished children to be optimized by targeting them to the specific needs of each child or region.

To implement such a strategy, a comprehensive framework must be developed to characterize the interplay of endogenous and exogenous metabolism in the context of acute malnutrition. This complexity raises the question of how best to characterize the microbiome among children with SAM relative to well-nourished children. With a few exceptions, previous microbiome studies typically focus on individual components of the gut ecology (e.g. operational taxonomic units [OTU]) and describe microbiomes in terms of taxa abundance and diversity. Although these metrics are informative, they overlook the fundamental interactions between the microbial and host ecosystems. With the prospect of targeted probiotics and interventional therapeutics on the horizon, the introduction of a single or a few bacterial species may be insufficient for the long-term nutritional restoration of a dysbiotic ecosystem.

Microbiota-directed therapeutics are relatively new in the realm of medicine, but the fundamentals and conceptualization have been pioneered extensively by Gordon and colleagues [24-27]. The hypothesis of undernutrition pathogenesis proposed in Blanton et al. 2016 is the following: 1) initial gut community is disrupted by one or more factors (e.g. enteropathogen competition); 2) gut dysbiosis provides opportunities for the virulence potential of pathobionts to be activated and enteropathogens to further establish themselves in the community and exacerbating dysbiosis; 3) impaired microbiome development impairs host immune response and defense; and 4) disruption of microbiota and host co-development induces the effects of undernutrition such as impaired muscle/bone growth, metabolism, and immune/gut barrier function.

As similar conditions of acute malnutrition can develop from different host-microbiome system configurations, the need for investigating the multifactorial nature of undernutrition and different mechanisms of pathogenesis is essential. The aim of our study was to investigate the interactions between the gut microbiome, enteric pathogens, and energy regulating hormones among rural Gambian children with non-edematous SAM (the predominant type of SAM in this region of The Gambia) during outpatient nutritional rehabilitation and MAM relative to well-nourished [WN] children. Our research leverages information gained from a sample-specific perspective towards providing insight into various mechanisms of pathogenesis and defining characteristics of the overarching nutritional status phenotypes. We applied innovative tools for analyzing these complex and dynamic interactions within an individual child.

2. Methods

2.1. Setting and context

The study was conducted in the East, West, and Central Kiang districts of rural Gambia. Rural Gambian infants are typically small at birth relative to international standards, show positive growth during the first few months of life, and then enter a period of profound growth faltering of both weight and length until 24 months of age [8]. Although breastfeeding up to 2 years of age is the norm, gastrointestinal infections are commonly acquired from as early as 3 months of age [28-30]. Most of the population are subsistence farmers and in recent years crop failure has been common, rendering them food insecure [31]. Housing conditions are often basic with animals living in close proximity to the humans. However, there has been increasing access to clean water and sanitation over the past 4 decades in some of the villages, although housing standards remain basic [32]. Most households rely on water from communal taps in the village and household level pit latrines for sanitation.

2.2. Study design and sample size

This was a cross-sectional observational sub-study of a quasi-experimental study that aimed to investigate the role of energy regulating hormones in the variable growth responses of rural Gambian children during nutritional rehabilitation. The baseline characteristics of the participants are published elsewhere [33]. In summary, children from 6 to 24 months of age who had presented to three rural primary health care services including the Medical Research Council Unit, The Gambia [MRCG] Keneba field station clinic, Soma Health Centre or villages covered by the Kwinella trekking team, from June 2013 to October 2014 were recruited. All participants recruited into the study underwent clinical assessments and assigned to one for the following groups: MAM, SAM, and well-nourished [WN (WHZ > -2)]; WN control participants were based on the anthropometric measurements [3]. MAM is classified as WHZ between 2 and 3 standard deviations below the WHO growth reference standard or a mid-upper arm circumference [MUAC] between 115 and 125 mm while SAM is classified as WHZ less than 3 standard deviations below or MUAC < 115 mm or bilateral oedema [3]. SAM children were managed in an outpatient nutrition rehabilitation unit in rural Gambia with a service for ambulatory provision of intravenous antibiotics and limited capacity for overnight observation and management of children requiring intravenous fluids and nasogatric feeds for < 48 hours. Children were who HIV-infected or had significant medical complications requiring transfer secondary or tertiary level care were excluded.

All children with SAM and MAM received 28 days of ready to use therapeutic foods (RUTF) according to the WHO and Gambian guidelines for the integrated management of acute malnutrition [34]. All children with SAM received broad spectrum antibiotics including amoxycillin for 7 days according to international guidance for management of SAM. In addition, 11 (61%) controls and 9 (41%) children with MAM had antibiotics prescribed at recruitment. Pre-and 1-hour post prandial venous blood samples for analysis of energy-regulating hormones were collected from all children at recruitment and for children in the MAM and SAM groups at days 14 and 28. Stool samples were collected from all the children at recruitment and at follow up visits at days 14 and 28 for children that presented with MAM or SAM (tables 2,3). Terminology, abbreviations, and methodology sources are available in table 1.

2.3. Ethics approval

Ethical approval to conduct this study was granted by the *Joint Medical Research Council Unit The Gambia* and *Gambia Government Ethics Committee*, L2015.14. All children were recruited into the study with written informed parental consent.

2.4. Clinical measurements for energy regulating hormones and receptors

The hormones were measured in plasma and blood as we described previously [33]. We calculated molar excess soluble leptin receptor [sOB-R] concentrations divided by leptin concentrations multiplied by 0.13 (i.e. $\frac{sOB-R}{Leptin} \times 0.13$), referred to as *molar* (table 4), a formula following Stein et al. 2006 in their study that sought to elucidate the role of the sOB-R and its regulation in children with protein energy malnutrition during nutritional recovery [35].

2.5. DNA extraction and enteric pathogen detection

Stool specimens underwent pre-treatment including glass bead beating followed by nucleic acid extraction using the QIAamp® Fast DNA stool mini Kit (Qiagen, Manchester UK). Efficiency of the nucleic acid extraction and amplification was monitored through external controls, bacteriophage MS2 and phocine herpesvirus as previously described[36]. Detection of 19 enteric pathogens (bacteria, viruses, helminths and protozoa) was performed using a custom TaqMan Array Card [TAC] previously developed to investigate the etiology of moderate-to-severe diarrhea among children[36]. TAC assays were run on the QuantStudio™ 7 Pro Real-Time PCR System (ThermoFisher Scientific, Loughborough, UK). The detailed procedures for setting up the TAC and cycling condition have been described elsewhere [36]. The gene target for *Escherichia/Shigella* in the custom Enteric TAC is the invasion plasmid antigen H (*ipaH)* which is carried by *Shigella* species and enteroinvasive *E. coli* (EIEC) [36].

Detections were considered negative if the cycle thresholds [ct] were greater than 35. Results were validated and discarded if any of the controls failed. The following nested pathogen markers were merged to avoid interdependence of variables: (1) *adenovirus* (*adenovirus_f, adenovirus_pan*); (2) *cryptosporidium* (*cryptosporidium, cryptosporidium_hominis, cryptosporidium_parvum*); (3) *giardia* (*giardia, giardia_a, giardia_b*), *norovirus* (*norovirusgi, norovirusgii*); and (4) *stec* (*stec_stx1, stec_stx2*). Pathogens that were detected in fewer than 3 samples were removed prior to downstream analysis including: *rotavirus_g1, vcholerae, cdifficile, cyclospora*, and *aeromonas*. As *epec_bfpa* and *epec_eae* are different gene targets on the *E. coli* adherence factor (EAF) plasmid and EPEC chromosome [37], respectively, we kept these separate to account for potential detection bias and mixed EPEC communities within a sample.

3. Quantification and statistical analysis

3.1. 16S rRNA, sequencing and operational taxonomic units

16S ribosomal RNA [rRNA] gene analysis was used to investigate the bacterial component of the gut microbiome. The 16S rRNA gene libraries (regions V3 and V4) were prepared using the Nextera® XT Index Kit (Illumina, Essex, UK) followed by multiplexed sequencing using the MiSeq Reagent Kit on the Illumina MiSeq System (Illumina, Essex, UK) following the manufacturer's protocol. Extraction and PCR controls were also included to detect possible contaminants. Operational taxonomic units [OTU] were generated *de novo* from Illumina sequence reads using *UPARSE* [38] and *mothur* [39] open-source bioinformatics tools. Paired-end reads were trimmed of adaptor sequences, barcodes, and primers prior to assembly, followed by discarding low quality reads and singletons. After a de-replication step and abundance determination, sequences were filtered for chimeras and clustered into OTUs. *UPARSE* has a built-in filter for chimera detection and removal, *UCHIME2* (http://www.biorxiv. org/content/early/2016/09/09/074252), which uses the highly curated SILVA database. To

Table 1
Terminology and abbreviations, Terminology and abbreviations for various concepts, data structures, and techniques used in this study.

Category	Name	Acronym	Description	Citation
Data	**Feature**		An individual measurable property or characteristic of a phenomenon being observed	
	Operational taxonomic unit	OTU	Groups of highly similar 16S rRNA sequences	
	Modality		One of the following datasets: (1) gut microbiome; (2) clinical measurements; and (3) pathogen markers	
	Multimodal		Data generated from multiple modalities or measurements	
Phenotypes	**Weight-for-Height Z-score**	WHZ	Compares a child's weight to the weight of a child of the same height and sex to classify nutritional status	[3]
	Well-nourished	WN	$-2 <$ WHZ	[3]
	Moderate acute malnourished	MAM	$-3 <$ WHZ ≤ -2	[3]
	Severe acute malnourished	SAM	WHZ ≤ -3	[3]
	Undernourished	UN	MAM or SAM. Used in the context aggregate networks and predictive modeling	
	Protein energy malnutrition	PEM	Protein-energy malnutrition defined as a range of pathological conditions arising from inadequate calories and/or protein intake.	[1]
Modeling	**Clairvoyance**		Feature selection algorithm leveraged for phenotype-discriminative community detection	[58]
	Hierarchical Ensemble of Classifiers	HEC	Graphical model where each internal node is a customized sub-model classifier with a unique feature set	[58]
	Sub-model		Machine-learning classification model used as internal node in a HEC model	
	Leave subject out cross-validation	LSOCV	Cross-validation designed to simulate performance on a new subject	This study
Networks	**Background network**	BN	Networks created from individuals who were WN for all	This study
	Perturbed background network	PBN	Networks created when adding in a query individual to the background network	This study
	Sample-specific network	SSN	Network with unique properties for each sample	[80]
	Sample-specific perturbation network	SSPN	Network created from perturbation between BN and SSPN distributions	This study
	Aggregate network	AN	Networks created from fitted sub-model coefficients	This study
	Node		The discrete objects within a network	
	Edge		Weighted connections between nodes	
	Edge weight		Association or perturbation strength of edge	
	Perturbation		Change in association strength of an edge between SSN and BN distributions	
	Connectivity	k	Sum of weighted edges connected to a note	
	Scaled connectivity	k∼	Scaled connectivity so total connectivity of all nodes sums to 1	
Pathogenic E. coli	**Enteropathogenic E. coli**	EPEC	EPEC E. coli are defined by the induction of a distinctive histopathology known as the attaching and effacing (A/E) lesion, which is characterized by the effacement of the intestinal microvilli and the intimate attachment of the bacteria to the host epithelial surface	[81,82]
	Enteroaggressive E. coli	EAEC	EAEC E. coli is defined as a diarrheal pathogen based on its characteristic aggregative adherence (AA) to HEp-2 cells in culture and its biofilm formation on the intestinal mucosa with a "stacked-brick" adherence phenotype.	[83]
	Enterotoxigenic E. coli	ETEC	ETEC E. coli are a pathogenic variant or pathovar of E. coli defined by production of diarrheagenic heat-labile (LT) and heat-stable (ST) enterotoxins.	[84]

predict taxonomy, we used the Wang classifier, and bootstrapped using 100 iterations. We set *mothur* to report full taxonomies only for sequences where 80 or more of the 100 iterations were identical (cutoff = 80). Taxonomies were assigned to the OTUs with *mothur* using version SSU Ref NR 99 123 of the SILVA 16S ribosomal RNA database as the reference. Tables with OTUs and the corresponding taxonomy assignments were generated and used in subsequent analyses with annotations detailed in table S2.

3.2. Alpha and beta diversity for microbial composition

Our preferred alpha diversity metric is microbial richness which measures the number of unique components detected within a particular sample. We measured alpha diversity for various levels including Family, Genus, Species, and OTU (figure 1 and supplementary figure 2). For all downstream OTU analysis, we removed OTUs that were not observed in at least 12% of the samples resulting in 155 prevalent OTUs. As OTU counts are compositional, we used a compositionally valid approach with the following protocol: (1) inferring a phylogenetic tree from OTU centroids using *FastTree v2.1.10* [40] with default parameters; (2) transforming abundances with the isometric log ratio transform using phylogenetic tree as the basis [41-43]; and (3) computing pairwise Euclidean distance [44].

3.3. Predictive functional profiling of microbial communities

We used *PICRUSt2 v2.3.0_b* to predict functional profiles for each sample [45] as human microbiomes are modeled with high performance [46]. We tested for differential enzyme abundance for MAM and SAM nutritional status phenotypes, with WN as reference, using the Mann-Whitney U test followed by Benjamini-Hochberg multiple hypothesis testing for adjusted p-values; statistical significance threshold set at 0.05. Statistical tests were run only for predicted enzymes that had more than 20 non-zero abundances in both WN and MAM/SAM classes as suggested by *SciPy* documentation. We used the GSEA's *Prerank* module (via *GSEApy v0.9.8*) to assess if any KEGG pathways were enriched in differentially abundant enzyme sets using 1 − adjusted p-values as our pre-ranked enzyme weight [47]. For *Prerank*, we used FDR < 0.25 for statistical significance as recommended by GSEA authors.

3.4. Modality-specific differential abundance analysis

We calculated differential abundance for each modality separately using appropriate methods for each data type. For our compositional microbiome data, we calculated differential abundance between WN/MAM and WN/SAM with the *ALDEx2 v1.20.0* R package using a Wilcoxon signed-rank test with Benjamini-Hochberg corrected p-values [48]. Our clinical measurements were non-integer continuous data and we calculated differential abundance using a Kruskal-Wallis H-test with normalized abundances [44]. Our pathogenic screening data was binary, so we used Fisher's exact test with a contingency table populated from the number of detected events and non-detected events for each nutritional status category. All p-values were corrected for multiple tests using Benjamini-Hochberg adjustment with an adjusted p-value threshold of 0.05 for statistical significance.

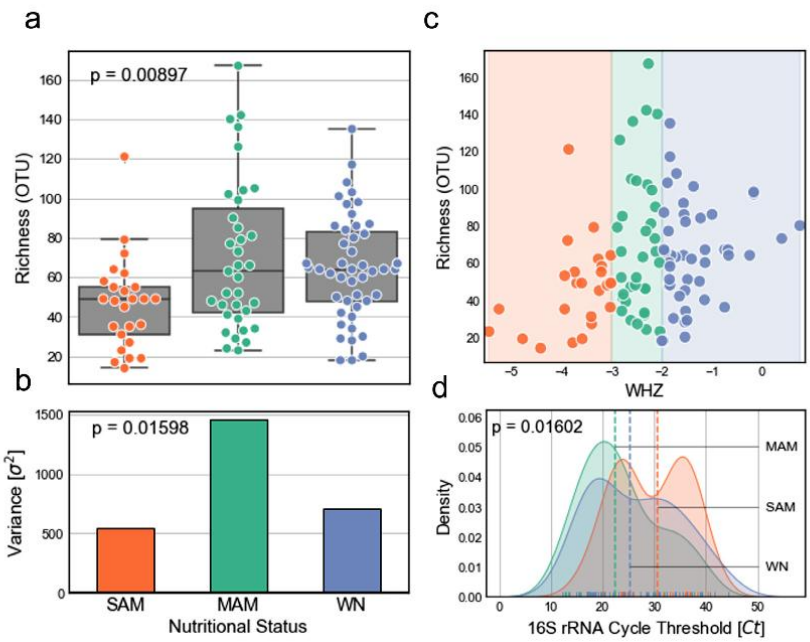


Figure 1. Microbial alpha diversity in the context of WHZ classifications, (a) Box-plot showing number of detected OTUs (richness) for each phenotype with inner values as interquartile range (Kruskal-Wallis H-test, p = 0.024). (b) Variance of microbial richness for each phenotype (Levene test, p = 0.019). (c) Microbial richness with respect to sample WHZ scores. (d) Cycle thresholds for 16S rRNA qPCR (Kruskal-Wallis H-test, p = 0.0169).

3.5. Linear mixed-effects regression

We used mixed linear-effects models to regress out variation from uncontrollable variables associated with the participant. In particular, we used the following regression model: $Feature_j \sim$ FixedEffects (*Age* + *Height* + *Sex*) + RandomEffect(*Participant Identifier*) where $Feature_j$ represents a normalized/transformed feature vector. *Age* and *Height* were z-score normalized prior to modeling. Features that statistically covary with microbial abundance, clinical measurements, and enteric pathogen presence are available in supplementary figure 4. We adjusted for these fixed and random effects in all subsequent analyses. We implemented multiple linear regression models using the *MixedEffectsLinearModel* object in *Soothsayer* [49] with the *statsmodels* [50] backend.

3.6. Network structure and visualization

Networks are graphical structures used to represent relationships between discrete objects where these discrete objects are referred to as nodes and connections between nodes as edges [51]. The connections are weighted by a numeric value.

The nodes represent individual features from the following modalities: (1) fecal gut microbiome; (2) clinical measurements; and (3) pathogenicity markers, while the edges between nodes represent intra- and inter-modality associations. To prepare the modalities for multimodal pairwise associations, we scaled the clinical measurements using z-score normalization and center log-ratio transformation for the fecal gut microbiome. As our multimodal data were only partially compositional, including an addition of binary and non-integer continuous data, we were not able to use proportionality metrics [52,53] and implemented a bootstrapped Pearson's correlation as our association metric.

Traditional ball-and-stick networks were visualized using the *NetworkX* Python package [54]. Hive plots were implemented using the *HiveNetworkX* Python package [55,56]. Edge and node connectivity RainCloud plots were generated using the *PtitPrince* Python package [57].

3.7. Sample-specific perturbation networks

Our objective with a sample-specific approach was to maximize the available data and quantify the amount by which a particular sample perturbs a biologically-relevant background distribution; we refer to this as a sample-specific perturbation network [SSPN]. We define a perturbation in the context of SSPNs as a change in association strength of an edge between background network [BN] and perturbed background network [PBN] distributions (supplementary figure 5). For the background cohort, we used data from participants whose nutritional status was WN for all visits (n=25, table 3), which allowed us to calculate SSPNs for participants that deviated to or from WN between visits (n=82 samples, table 2).

More specifically, we created SSPNs for 82 samples using 1000 sub-sampled permutations and a sampling size [$\boldsymbol{n}$] of 15 with each draw denoted as $\boldsymbol{k}$. The ***BN*** distribution was created by pseudo-randomly (seed=0) drawing $\boldsymbol{n}$ samples from the background cohort without repetition, creating a pairwise multimodal association network, and stacking the edges for all permutations allowing us to calculate a distribution of background edge weights; ***BN*** is parameterized by a matrix of $\boldsymbol{k}$ draws and $\boldsymbol{m}$ edges. For each query sample $\boldsymbol{i}$, we repeated the ***BN*** creation process using the same sub-sampled sample draws but we added the query sample to build a distribution of edges for our query PBN distribution [$\boldsymbol{PBN_i}$]. For each edge $\boldsymbol{j}$, we calculate the difference in means for the query network distribution $\boldsymbol{PBN_{ij}}$ and background network distribution $\boldsymbol{BN_j}$ to build

Table 2
Baseline characteristics, Baseline characteristics of children with respect to nutritional status as defined in a previous study [33].

	Nutritional category			P
	WN (N=22)	MAM (N=18)	SAM (N=20)	
Age in months, median (IQR)	12.75 (10.2, 19.3)	16.5 (12.0, 22.0)	12.0 (10.3, 16.5)	0.22[a]
Age of weaning in months, median (IQR)	6.0 (5.0, 6.0)	6.0 (6.0, 6.0)	6.0 (5.5, 6.0)	0.82[a]
WHZ, median (IQR)	-1.2 (-1.8, 0.1)	-2.6 (-2.8, -2.1)	-3.4 (-3.9, -3.2)	
WAZ, median (IQR)	-1.5 (-1.7, -0.1)	-2.8 (-3.1, -2.1)	-3.2 (-3.4, -2.9)	<0.001[a]
HAZ, median (IQR)	-0.7 (-1.8, 0.03)	-1.7 (-2.5, -1.0)	-1.9 (-2.3, -0.9)	0.08[a]
Salivary CRP, ng/mL, median (IQR)	2.9 (2.4, 4.1)	4.9 (2.8, 10.3)	5.6 (4.1, 9.9)	0.04[a]
Urinary tract infections, n (%)	4 (19)	2 (12)	1 (6)	0.47[b]
*Diarrhea, n (%)	2 (11)	4 (29)	11 (58)	0.01[b]
**Antibiotics prescribed	9 (41)	11 (61)	18 (90)	0.003[b]
Females, n (%)	11 (50)	8 (44)	10 (50)	0.90[b]

Abbreviations: IQR Interquartile range, MAM Moderate Acute Malnutrition; SAM Severe Acute Malnutrition; WHZ weight-for-height z-score; WAZ weight-for-age z-score; HAZ, Height-for-age z-score.
[a] Kruskal-Wallis test.
[b] Fisher's exact test.
* Diarrhea defined according to the WHO guidance of passage of 3 or more loose or liquid stools per day or more frequent passage than normal for the child.
** If the child had any antibiotics prescribed during the study (irrespective of type, mode of administration or duration).

$\boldsymbol{SSPN_i}$(a vector for sample $\boldsymbol{i}$ with $\boldsymbol{m}$). Each edge $\boldsymbol{j}$ in $\boldsymbol{SSPN_i}$ is a measure of how much the addition of query sample $\boldsymbol{i}$ perturbs the background system. Although the BN and PBN distributions are intermediate data structures used solely to calculate SSPNs, these network distributions are fundamental to capturing true variability within a study cohort while also being robust to sample outliers. More algorithm details are available in the source code mentioned below.

The range of values for each SSPN edge weight is in the closed interval [-2,2] where positive and negative values indicate an increase and decrease in association strength, respectively. Each SSPN was stacked to yield a perturbation matrix [$\boldsymbol{M}$] of 82 samples and 14,270 edges, where each row $\boldsymbol{i}$ corresponds with a sample, each column $\boldsymbol{j}$ represents an edge, and each $\boldsymbol{M_{i,j}}$ indicates a perturbation value. This matrix structure allowed us to efficiently store networks and leverage machine-learning methods that require 2-dimensional data. These methods are open-sourced in our *EnsembleNetworkX* Python package using the *SampleSpecificPerturbationNetwork* object.

3.8. Phenotype-discriminative network community detection

We decomposed our classification of nutritional status phenotypes into step-wise binary classifications of severity using highly interpretable algorithms for each step. Each step in our hierarchical classification scheme can be interpreted which maximizes the available information content.

For our training data, we use the perturbation matrix $\boldsymbol{M}$, containing our SSPNs, as our feature matrix and their associated nutritional status classifications as the target vector. To ensure our associations were not biased, and remained clinically relevant, we used a stringent Leave Subject Out Cross-Validation [LSOCV] to simulate classification performance on new participants and LSOCV accuracy as a proxy for reliability in clinical relevance. Although our primary objective was not to classify sample nutritional status based on SSPNs, we used LSOCV to simulate classification performance on new participants and LSOCV accuracy as a proxy for the community detection capabilities of our feature selection analysis.

In particular, we implemented a Hierarchical Ensemble of Classifiers [HEC] model to partition a single tertiary classification into two step-wise binary classifications. Our HEC model asks the following questions: (1) is the participant at this visit WN or acutely malnourished and (2) if undernourished, is the participant at this visit MAM or SAM? We translate these human interpretable questions directly to machine interpretable tasks by building a step-wise classification algorithm that trains each decision using a subset of the samples and a subset of features. In particular, we produce the following HEC model: 1) the first binary classification performed by sub-model *y1* discriminates between WN and MAM/SAM, collectively referred to as undernourished [UN] class for here forth; and (2) the second binary classification performed by sub-model *y2* differentiates between MAM and SAM. The order of each binary class mentioned prior corresponds with a 0 and 1 in a standard logistic regression classification. In these fitted logistic regression models, a positive coefficient corresponds with an increase in likelihood that a sample is classified as the 2nd class (i.e. UN in sub-model y1 and SAM in sub-model y2) and vice versa. Despite being the logical progression when diagnosing severity, these human interpretable questions that define the step-wise decisions in our HEC model were data-driven and determined by *Soothsayer's Topology* method using only the training data.

These sub-models were optimized using the *Clairvoyance* feature selection algorithm (available within the *Soothsayer* package) which returns a set of features, edges in this case, and sub-model hyperparameters that result in the corresponding LSOCV accuracy [49,58]. We used feature selection with our SSPNs to identify edges that are associated with discriminating phenotypes (i.e. paths within the graph). The output of the feature selection algorithm includes 3 main elements including the following: (A) hyperparameters for the sub-model classifier; (B) the edge set used during model fitting; and (C) the LSOCV accuracy using a combination of using (A) hyperparameters with (C) features. We selected the set of edges and hyperparameters with the highest LSOCV accuracy to build classification sub-models as internal nodes in the HEC model using *Soothsayer's HierarchicalClassifier* method. Edge sets $\boldsymbol{Edges_{y1}}$ and $\boldsymbol{Edges_{y2}}$ were used to build sub-models *y1* and *y2*, respectively, representing the smallest subset of features that most effectively discriminate between nutritional status phenotypes. *Clairvoyance* feature selection and LSOCV accuracy have been adapted from Espinoza and Dupont et al. 2021 which were developed to model antimicrobial mechanism-of-action [58].

We used L2-regularized logistic regression classifiers for both sub-models ***y1*** and ***y2*** with inverse regularization strength of 1.0 and 0.106, respectively, as these machine-learning algorithms often perform with high accuracy and are human interpretable compared to "black-box" algorithms such as neural networks [59]. Logistic regression classification models are easily interpreted as each feature has a coefficient and the magnitude of these coefficients reflects the influence a particular feature has on the classification.

We used *Clairvoyance* with the following parameters: *–model_type logistic,tree –n_iter 500 –min_threshold None, 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 –percentiles 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.91, 0.92, 0.93, 0.94, 0.95, 0.96, 0.97, 0.98, 0.99 –method bruteforce –early_stopping 100 –cv LSOCV.tsv* where LSOCV.tsv contains custom training and testing pairs organized by participant.

3.9. Aggregate networks

We developed ANs as a method to quantify the predictive capacity of a node, edge, or subgraph in the context of discriminating phenotypes. ANs serve as an edge framework for building phenotype-discriminative SSPNs whose weights are populated by sample-specific perturbations from $\boldsymbol{M}$. In the case of this study, AN edge weights represent L2-regularized logistic regression coefficients; though, it would be seamless to incorporate L1-regularization or feature importance metrics from tree-based algorithms. More specifically, $\boldsymbol{Edges_{y1}}$ and $\boldsymbol{Edges_{y2}}$ were used to build $\boldsymbol{AN_{y1}}$and $\boldsymbol{AN_{y2}}$with edges weighted by the fitted ***y1*** and ***y2*** sub-model parameters. A positive coefficient in sub-model ***y1*** and ***y2*** corresponds with an increased likelihood for classifying a sample as UN or SAM, respectively, where UN represents either MAM or SAM. The inverse of this is true as well where a negative coefficient represents a decrease in likelihood in said classes. Aggregate and sample-specific perturbation networks were implemented using *Graph* objects in *NetworkX* [54] and *EnsembleNetworkX*.

3.10. Connectivity in the context of sample-specific networks and aggregate networks

Network connectivity [k] is a metric used to quantify the influence a particular edge, node, or group of nodes has within a system. In this study, we implemented weighted-degree as our connectivity metric which can be measured at different levels. We define connectivity at the edge level to represent the edge weight of a pair of nodes, while at the node level connectivity refers to the sum of weighted edges connected to a node. These connectivities can be grouped such as quantifying the total connectivity of a subset of edges (as in the case for node connectivity by grouping edges connected to a node), a subset of nodes, or the entire network itself. Scaled connectivity [$k^{\sim}$] normalizes the node connectivity values so the total connectivity within a network sum to 1 and can be used to compare networks with different numbers of nodes or edges; such is the case for comparing ANs.

Network connectivity is interpreted differently depending on the network context. In the context of SSPNs, edge weight represents perturbation magnitude; that is, the change in association of a query PBN distribution with respect to a BN distribution. Therefore, the total connectivity within a SSPN measures how much a particular sample perturbs the background associations with respect to the edges in the network. In the context of ANs, L2-regularized logistic regression coefficients are used as edge weights and can be interpreted as the influence of an edge in predicting nutritional status phenotypes. Therefore, the total connectivity within ANs is the combined influence of the edges in their predictive capacity.

3.11. Recovery scores

Our dataset contains time-ordered samples for many of the participants. We developed an edge recovery score [$\boldsymbol{r}$] to quantify the amount in which an edge contributes to weight recovery; more specifically, the transition from UN to WN. As SSPN edge weight indicates a perturbation between the PBN and BN distributions, we identified perturbations relevant to nutritional status recovery by selecting for edges in consecutive time-ordered SSPNs that have the following properties: (1) greatest change in perturbation of associations between visits t_n and t_{n+1}; and (2) smallest perturbation magnitude from the BN distribution at t_{n+1} where t_n and t_{n+1} phenotypes represent UN and WN, respectively.

For each participant that recovers their nutritional status, we calculate $\boldsymbol{r}$ using the following equation: $r_{i,j}(x,y) = \frac{|y-x|^2}{4} + 6|y|$ where x and y represent edge weight j for participant i at visits t_n and t_{n+1}, respectively. The expression $|y-x|^2$ corresponds with property (1) while $4 + 6|y|$ corresponds with property (2). This equation bounds $\boldsymbol{r}$ within the closed interval [0,1] where higher values correspond to larger potential influence in weight recovery (figure 6a). Non-zero recovery scores were grouped by (1) MAM → WN and (2) SAM → WN transitions between t_n and t_{n+1}. These distributions were plotted using RainCloud plots and outlier thresholds were determined using the 1.5*IQR + Q3 between the distributions where IQR is interquartile range and Q3 is the 75th percentile. The outlier thresholds for MAM → WN (0.449 r) and SAM → WN (0.452 r) recovery scores were very similar so we used the minimum value of 0.449 r as our consensus threshold for outliers (figure 6b).

The recovery score metric condensed the information content of complex multimodal time-ordered SSPNs into a single human interpretable metric. We designed the recovery score to demonstrate the following properties for each edge: (1) a large difference between an undernourished visit and a consecutive WN visit; and (2) a small edge weight for WN. Emphasizing these properties allowed us to collapse the temporal dimension, the sign of edge weights, and focus specifically on edges specifically to the recovery of an individual participant.

3.12. Role of funding source

The funding sources had no role in the design of this study and did not have any role in the study design, data collection, data analyses, interpretation, or writing of report. or decision to submit results.

4. Results

4.1. Dataset overview

Sixty children were recruited to the study (20 SAM, 18 MAM, 22 WN) with visit-specific samples detailed in tables 2,3. All the children with SAM presented with non-edematous SAM. The median age of the children was 12.0 months for SAM, 16.5 months for MAM, and 12.75 months for WN at recruitment. A total of 54 children had stool specimens available in this study that passed quality control; 25, 35 and 47 stool samples were collected from the children in SAM, MAM and WN groups based on their nutritional status at baseline. Stool samples were collected only at baseline among the WN children. For the children in MAM and SAM, stool samples were collected at baseline, Day 14 and Day 28 (table 3 and supplementary figure 2e).

Following nutritional intervention, 36 participants with acute malnutrition (i.e. MAM and SAM) were included in follow up visits. Of these follow up visits, 16 participants maintained a consistent nutritional status throughout all visits (SAM: 3 participants, MAM: 8 participants, and WN: 5 participants). Sixteen participants showed signs of recovery of at least one nutritional status level (i.e. SAM → MAM, SAM → WN, or MAM → WN), with 13 participants recovering between $t_0 \rightarrow t_{14}$ and 3 participants between $t_{14} \rightarrow t_{28}$ visits (supplementary figure 2e). Four participants showed signs of decline of at least one nutritional status level (i.e. WN → MAM, WN → SAM, or MAM → SAM) between $t_{14} \rightarrow t_{28}$ visits. Multiple modalities were measured including: 1) fecal microbiome 16S rRNA sequencing (388 OTUs); 2) 12 clinical measurements related to immune activation, inflammation, and energy regulating hormones, and 3) qPCR derived presence of 23 viral, fungal, protozoan and bacterial enteric pathogens.

Table 3
Subjects and samples with respect to nutritional status, The number of subjects and samples with respect to nutritional status and status changes.

Visits	Category	Phenotype	N Participants	N Samples
All visits	**Nutritional status**	SAM	16	25
		MAM	25	35
		WN	34	47
		Total	75 (54 unique)	107
	Maintained nutritional status	SAM	4	10
		MAM	10	19
		WN	20	25
Follow-up visits	**Nutritional status**	SAM	15	24
		MAM	23	33
		WN	19	32
		Total	57 (36 unique)	89
	Maintained nutritional status	SAM	3	9
		MAM	8	17
		WN	5	10
	Recovery	t=0 → t=14		
		SAM → MAM	3	-
		SAM → WN	5	-
		MAM → WN	5	-
		Total	13	
		t=14 → t=28		
		SAM → MAM	2	-
		SAM → WN	0	-
		MAM → WN	1	-
		Total	3	-
	Decline	t=0 → t=14		
		WN → MAM	0	-
		WN → SAM	0	-
		MAM → SAM	0	-
		Total	0	-
		t=14 → t=28		
		WN → MAM	3	-
		WN → SAM	0	-
		MAM → SAM	1	-
		Total	4	-

4.2. Fecal gut microbial composition and functional profiling predictions

There were a total of 388 high quality OTUs that passed quality control. The number of 16S ribosomal (r)RNA gene reads mapped to OTUs ranged from 392 - 92,910 and the number of detected OTUs (richness) ranged from 14 - 167 per sample (figure 1 and supplementary figures 1,2a-d). The microbial richness of SAM samples was statistically lower than both WN and MAM samples (figure 1a,c; Kruskal-Wallis H-test, p = 0.009). MAM samples had statistically higher variance compared to WN and SAM (figure 1b; Levene test, p = 0.016). The variance in richness was the highest for WHZ in the range (-3,-1), the entirety of MAM samples and the lower WHZ of WN samples (figure 1c).

We observed evidence of differences in 16S rRNA qPCR *Ct* values between nutritional status phenotypes (figure 1d; Kruskal-Wallis H-test, p = 0.016). MAM samples exhibited the lowest median *Ct* and thus the highest relative biomass on average (22.5 *Ct*), while SAM exhibited the lowest (30.5 *Ct*) with WN in between (25.4 *Ct*). A bimodal distribution of SAM *Ct* values with the lower peak at ~24 *Ct* and the upper at ~35 *Ct* suggested that while SAM samples have similar community composition (figure 2), in a substantial number of cases the bacterial community had collapsed from a numerical perspective.

The relative abundances of bacterial phyla were very similar across all nutritional status phenotypes with a few notable exceptions (figure 2a). The combination of *Firmicutes* (WN = 46%, MAM = 50%, SAM = 33%), Bacteroidetes (WN = 28%, MAM = 20%, SAM = 17%), and Proteobacteria (WN = 20%, MAM = 23%, SAM = 45%) constitute more than 90% of the microbial abundance regardless of the participant's nutritional status with respect to visit. We did not observe any components at any level of taxonomy that were differentially abundant among children with MAM relative to WN. We did however observe an enrichment in *Enterobacteriaceae* abundance (Wilcoxon signed-rank test, adjusted p = 0.018) for SAM (μ = 42%) relative to MAM (μ = 19%) and WN (μ = 18). The only differentially abundant OTU was an unclassified *Klebsiella* (*Otu000014*) with an enrichment in SAM (μ = 16%) relative to MAM (μ = 6%) and WN (μ = 3%) (adjusted p = 0.0442).

Beta diversity analyses did not reveal any defining global patterns associated with WN, MAM, or SAM (figure 2b,c). Despite this lack of qualitative separation in ordination space and through hierarchical clustering, we found evidence of differences between intra-and inter-nutritional category beta diversity (figure 2d; Mann-Whitney rank test, $p < 0.001$). The beta diversity of microbial communities from the same participants was lower than within a nutritional category (figure 2e; $p < 0.001$), despite some participants transitioning across nutritional categories during follow-up.

Functional content predictions of the gut microbiome were performed to gain insight into potential metabolic characteristics of each nutritional status phenotype. Predictive functional profiles produced 1390 predicted enzymes that were inferred from the fecal microbiome There were not any differentially abundant predicted enzymes relative to WN in MAM but 81 enzymes were differentially abundant in SAM. This set of 81 enzymes were enriched in 4 KEGG pathways including *map01120* (Microbial metabolism in diverse environments), *map00360* (Phenylalanine metabolism), *map00362* (Benzoate degradation), and *map01100* (Metabolic pathways).

4.3. Clinical measurements related to energy-regulating hormone

We explored the differences in plasma levels of key child growth and energy-regulating hormones at baseline for all the children and

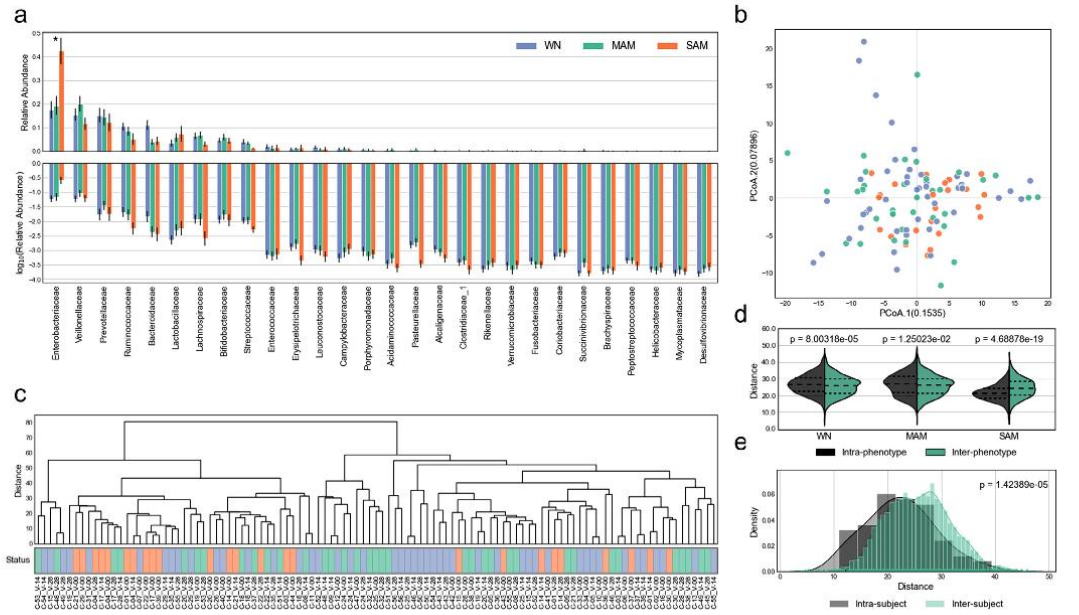


Figure 2. Taxonomic abundance and microbial beta diversity, (a) Relative abundance of OTUs summed by family-level taxonomy displayed using (A_{Top}) linear and (A_{Bottom}) log scales (Kruskal-Wallis H-test, adjusted p = 0.0246). (b) Principal coordinates analysis plot using phylogenetic isometric log-ratio transformation [PHILR] and Euclidean distance as precomputed distance matrix. (c) Hierarchical clustering using PHILR distance matrix (d,e) Intra- and inter-grouping diversity using PHILR distance matrix for (d) phenotype (Mann-Whitney rank test) and (e) participant (Mann-Whitney rank test, p = 1.424e-5). Error bars represent standard error of the mean.

at days 14 and 28 for children with MAM and SAM. The annotations and descriptions for clinical measurements are summarized in table 4. Several clinical measurements were differentially abundant between nutritional status categories. In particular, *leptin* (Kruskal-Wallis H-test, adjusted $p < 0.001$), IGF-1 [*igf1*] (adjusted $p < 0.002$), and IGF-binding protein-3 [*igfbp3*] (p = 0.003) increased with nutritional status in that WN have the highest levels (figure 3a). As these statistics are associated with nutritional status which are dependent on WHZ, correlations between normalized clinical measurements and WHZ scores showed that *leptin* (Pearson $p < 0.001$), IGF-1 ($p < 0.003$) and its binding protein IGFBP-3 ($p < 0.001$) were indeed positively correlated with WHZ. In contrast, the molar excess of soluble leptin receptor (sOBR)/leptin [*molar*] (adjusted $p < 0.001$), *ghrelin* (adjusted p = 0.002), ghrelin receptor [*ghrp*] (adjusted p = 0.003), and sOB-R [*sobr*] (adjusted $p < 0.007$) measurements decreased monotonically with nutritional status in that SAM have the highest levels (figure 3a). As expected, the molar excess of sOBR/leptin ($p < 0.001$), *ghrelin* (p = 0.012), *ghrp* (p = 0.012), and *sobr* (p = 0.011) levels were inversely correlated with WHZ.

4.4. Enteric pathogens and virulence factor prevalence

The abundance of 23 enteric pathogen markers was assessed using qPCR. All children had between 1 to 15 pathogen or virulence factors detected in at least one visit (figure 3b). In accordance with microbial richness, we observed MAM to have a greater number of pathogenic markers (8 pathogens) compared to SAM or WN (7 pathogens) nutritional status phenotypes. Several pathogenic markers were differentially prevalent between the nutritional status phenotypes. In particular, *Giardia duodenalis* was significantly less prevalent among children with SAM (23%) than MAM (68%) or WN (70%) (Fisher's exact test; adjusted $p = < 0.001$). In contrast, the prevalence of enteropathogenic *E. coli* with the *bfpA* (adjusted p = 0.007) and *eae* (adjusted p = 0.024) virulence increased with severity of malnutrition (figure 3b). Bundle-forming pilus A [*bfpA*] and intimin adherence protein [*eae*] are genes found on the EAF plasmid and EPEC genome, respectively, and contribute to attachment to epithelial cells, thus, leading to the attaching and effacing phenotype [60,61].

4.5. Identifying perturbations capable of discriminating nutritional status phenotypes

Our SSPN analysis produced a perturbation matrix (N= 82 samples, M= 14,270 edges representing 188 nodes) (supplementary figure 5). By analyzing our SSPNs using a HEC model and feature selection in unison, the most accurately predictive edges within the network were identified. Essentially, this characterizes which interactions are most predictive of nutritional phenotype and is more informative for nutritional intervention than regression methods predicting a rise or fall in WHZ and disregarding intra-phenotype patterns. By selecting for only the edges that are informative in discriminating nutritional status phenotypes, and by extension the informative nodes, the information content in the edges was compressed by 98.143%; that is, 265 of the 14,270 edges.

Unsupervised machine-learning can be used to gain insight into the underlying structure of the data and leveraged to validate dimensionality reduction (e.g. feature selection) results based on predefined categories, nutritional status in this context. Unsupervised clustering of held-out prediction probabilities (figure 4c) were more homogenous than unsupervised clustering based on microbiome abundance profiles (figure 2c) or pathogen markers (supplementary figure 3). Clustering by held-out prediction probabilities revealed that the edge features in each sub-model capture biologically relevant discriminatory patterns. It should be noted that the

Table 4
Clinical measurements and pathogenic markers, Description of clinical measurements and pathogenic markers used in study with respect to shortened feature label.

Feature	Modality	Description
sobr	clinical	Fasting soluble leptin receptor [sOB-R] concentration
leptin	clinical	Fasting leptin concentration
ghrp	clinical	Fasting ghrelin receptor [ghrelinR] concentration
ghrelin	clinical	Fasting ghrelin concentration
molar	clinical	Fasting molar excess sOB-R concentration/ leptin ratio
igf1	clinical	Fasting insulin-like growth factor 1 [IGF-1] concentration
igfbp3	clinical	Fasting insulin-like growth factor binding protein 3 [IGFBP-3] concentration
molarigf1igfbp3	clinical	Fasting molar excess of IGF-1 concentration/ IGFBP-3 ratio
insulin	clinical	Insulin concentration
salivarycrp	clinical	Salivary C-reactive protein concentration
cortisol	clinical	Cortisol concentration
astrovirus	pathogen	Astrovirus
bfragilis	pathogen	Bacteroides fragilis
campylobacter_pan	pathogen	Campylobacter pan-genomes
cryptosporidium	pathogen	Cryptosporidium
eaec_aaic	pathogen	Enteroaggregative Escherichia coli aaiC gene
eaec_aar	pathogen	Enteroaggregative Escherichia coli Aar gene
eaec_aata	pathogen	Enteroaggregative Escherichia coli aatA gene
eaec_aggr	pathogen	Enteroaggregative Escherichia coli AggR gene
ebieneusi	pathogen	Enterocytozoon bieneusi
enterovirus	pathogen	Enterovirus
epec_bfpa	pathogen	Enteropathogenic Escherichia coli bfpA gene
epec_eae	pathogen	Enteropathogenic Escherichia coli eae gene
etec_lt	pathogen	Enterotoxigenic Escherichia coli heat-labile enterotoxin gene
etec_sth	pathogen	Enterotoxigenic Escherichia coli heat-stable enterotoxin gene
etec_stp	pathogen	Enterotoxigenic Escherichia coli heat-stable enterotoxin gene
giardia	pathogen	Giardia duodenalis
hpylori	pathogen	Helicobacter pylori
salmonella	pathogen	Salmonella
sapovirus	pathogen	Sapovirus
shigellaeiec	pathogen	Shigella/Enteroinvasive Escherichia coli
stec	pathogen	Shiga toxin-producing Escherichia coli

unsupervised clustering in figure 4c used Euclidean distance of probabilities generated from multimodal edges while figure 2c and supplementary figure 3 used Euclidean distance of individual modalities. We were not able to implement unsupervised clustering of the multimodal nodes using Euclidean distances because the pathogenic markers are binary and clinical measurements had missing data.

4.6. Connectivity signatures in aggregate networks

We implemented aggregate networks [AN] from each sub-model to structure the edge perturbations that could accurately predict nutritional status phenotype (figure 5). Edge weight in the context of ANs corresponds with the capacity of a specific edge to predict nutritional status phenotype and these networks serve as a method to quantify the predictive capacity of specific nodes and edges for clinical interpretation. As the training data for the parent sub-models are edge perturbations with respect to a sample in a SSPN, an increase in association strength (a positive perturbation) and a positive coefficient indicate an increased likelihood for said classes (See *Methods*).

The higher scaled node connectivity (Mann-Whitney rank test, $p = < 0.001$) and lower edge connectivity (Mann-Whitney rank test, $p < 0.001$) in $\boldsymbol{AN_{y2}}$ revealed there were a few nodes with high predictive capacity compared to $\boldsymbol{AN_{y1}}$ which was closer to a normal distribution. In $\boldsymbol{AN_{y1}}$, we observed an unclassified *Escherichia-Shigella* (*Otu000281*), as the one of two nodes with substantially more influence in discriminating WN from malnourished children. The highest weighted edges to this unclassified *Escherichia-Shigella* were through *ghrelin* (-0.321 *k*) and *ghrp* (0.311 *k*) (supplementary figure 6a). The second node with substantially higher influence in $\boldsymbol{AN_{y1}}$ was *molar* (0.0384 $k^{\sim}$). The highest connectivity edges to *molar* were through *Lactobacillus mucosae*, an unclassified *Haemophilus* (0.289 *k*), and an unclassified *Ruminococcaceae* UCG-002 (0.264 *k*) (supplementary figure 6b). The only edge in AN_{y1} with substantially greater edge connectivity relative to the other edges in the network was *Otu000965*:unclassified *Pantoea* — *Otu000906:Bifidobacterium stellenboschense* (-0.483 *k*).

The scaled node connectivity in AN_{y2} was greater than AN_{y1} and this difference was largely influenced by two outlier *Enterobacteriaceae* nodes with much higher connectivity relative to the other nodes in the network: (1) an unclassified *Pantoea* (0.0519 $k^{\sim}$) and (2) an unclassified *Enterobacteriaceae* (0.0409 $k^{\sim}$). Therefore, these two nodes accumulated 9.27% of the total connectivity within AN_{y2}. When grouping nodes by family-level taxonomy, we observed that *Enterobacteriaceae* constituted 14.61% of the total connectivity in AN_{y2}. The greater edge connectivity in AN_{y2} was largely influenced by four high connectivity edges: (1) *Weissella cibaria* — *Lactobacillus oris* F0423 (-0.222 *k*); (2) *igf1* — unclassified *Enterobacteriaceae* (-0.220 *k*); (3) unclassified *Streptococcus* —unclassified *Megasphaera*, (-0.201 *k*); and (4) *Prevotella* — *Subdoligranulum* (0.191 *k*).

4.7. Quantifying changes in sample-specific perturbation networks over time

Time-ordered SSPNs provide unique insight into how these multimodal systems evolve over time in relation to nutritional status recovery. We developed the recovery score [$\boldsymbol{r}$] to reduce the information content of time-ordered SSPNs and rank edges by their potential contribution to nutritional status recovery. The recovery score is designed to reward edges with large differences between $\boldsymbol{t_n}$ and $\boldsymbol{t_{n+1}}$ while penalizing larger values of $\boldsymbol{t_{n+1}}$ where $\boldsymbol{t_n}$ and $\boldsymbol{t_{n+1}}$ represent edge weights for UN and WN consecutive visits, respectively. This property of the recovery score can be visualized with the convex shape defined by $\boldsymbol{r(x,y)}$ as lower scores exist when $\boldsymbol{t_n}$ and $\boldsymbol{t_{n+1}}$ are similar or when $\boldsymbol{t_{n+1}}$ is large (figure 6a).

There were a total of 67 edges representing 76 unique nodes from 12 participants that had recovery scores statistically greater than the IQR ($r_{threshold} \geq 0.449$). The 12 participants in this statistically significant group are represented by 6 MAM to WN transitions (MAM → WN) and 6 SAM to WN transitions (SAM → WN). The 76 nodes from edges with statistically significant recovery scores included 69 OTUs (40 *Firmicutes*, 12 *Proteobacteria*, 12 *Bacteriodetes*, 4 *Actinobacteria*, and 1 *Fusobacteria*), 3 clinical measurements (*leptin, molar, salivarycrp*), and 4 enteric pathogens (*ebieneusi, epec_bfpa, etec_stp, stec*). Unsupervised clustering of statistically significant recovery scores did not reveal any noticeable groupings by MAM → WN or SAM → WN transitions (figure 6c). However, one group of transitions (N=2 MAM → WN and N=1 SAM → WN) with participants aged from 20.3 – 24.4 months formed a tight cluster outside of the remaining 9 transitions. Both MAM samples in these transitions have WHZ ≤ -2.79, close to the -3 WHZ threshold for SAM, which suggests a similar nutritional status to the lone SAM → WN transition in the cluster.

Many outlier recovery scores were unique to a participant within our cohort. However, there were a few exceptions where we observed more consistency across participants with similar nutritional status shifts. In particular, 3 edges had statistically high recovery scores in multiple MAM → WN transitions including OTUs belonging to *Prevotellaceae, Lachnospiraceae, Phascolarctobacterium and Ruminococcaceae* species. We also observed 3 edges that had statistically high recovery scores in more than one SAM → WN transition including *Lachnospiraceae, Coriobacteriaceae, Prevotellaceae,*

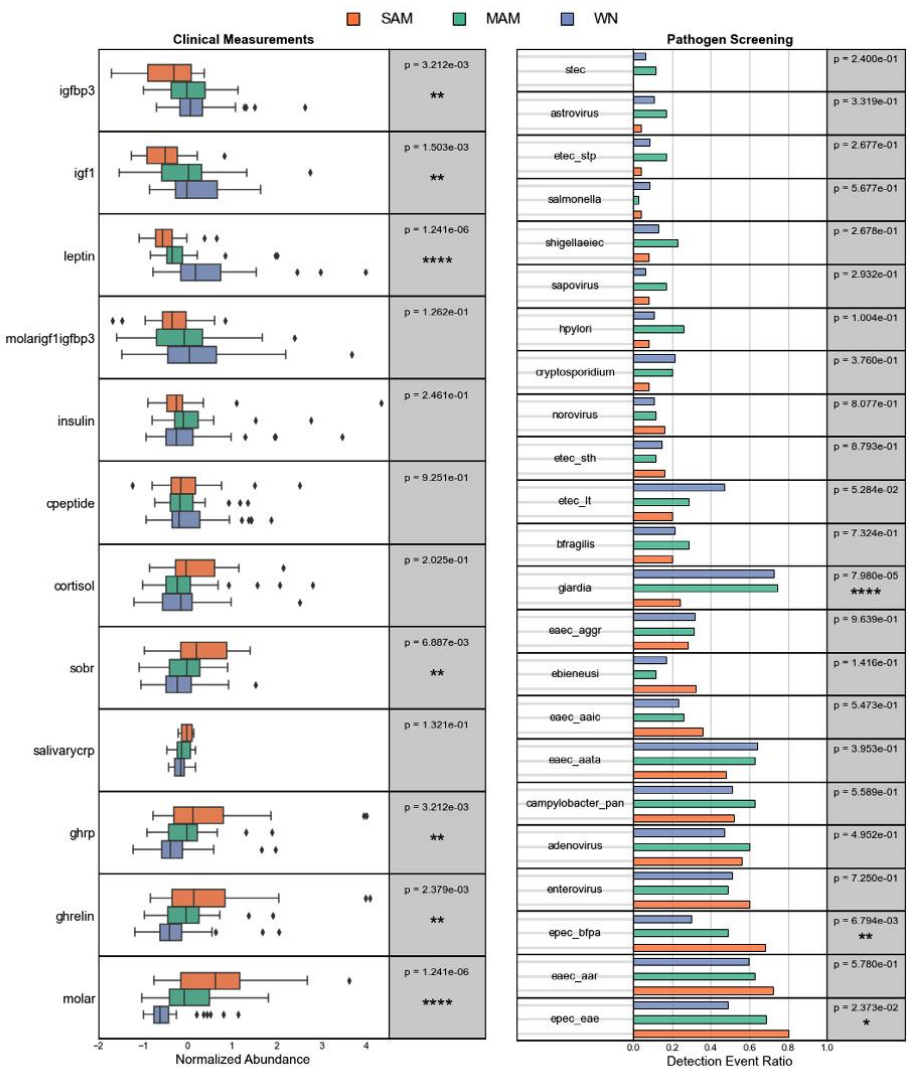


Figure 3. – Clinical measurement and pathogen abundance, (a) Boxplots showing distributions of normalized clinical measurements with respect to nutritional status with whiskers indicating IQR. Statistical test used was Kruskal-Wallis H-test. (b) Ratios of positive pathogen detection events with respect to nutritional status. Statistical test used was Fisher's exact test.

Subdoligranulum and Bacteroides and *Enterobacteriaceae* species. The cohort sample size was not large enough to statistically model specific associations in a participant's ability to decline or recover with respect to their nutritional status.

5. Discussion

In this study, we have shown that children with non-edematous SAM (marasmus) that is the more prevalent type in this region of The Gambia, have distinct microbiome characteristics and biologically-relevant multimodal biomarkers not observed in MAM and WN nutritional phenotypes. These findings also provide some key preliminary insights into systematic changes in gut microbiota and host energy regulating hormones among children with marasmus SAM during outpatient nutritional rehabilitation and among under 5's with marasmus SAM, MAM and WN nutritional phenotypes, that may have important implications for future research into prevention and treatment strategies of acute malnutrition [24].

Analyzing each data type (modality) independently allowed us to validate our findings by cross-referencing against previous studies and set the context for our multimodal network analysis. Consistent with previous reports [12,13,15,62], we found that children with SAM had significant reductions in richness and bacterial loads compared to WN or MAM participants. It is possible that the bacteria depleted in acutely malnourished children are essential for optimal digestion, nutrient absorption, modulating inflammation and

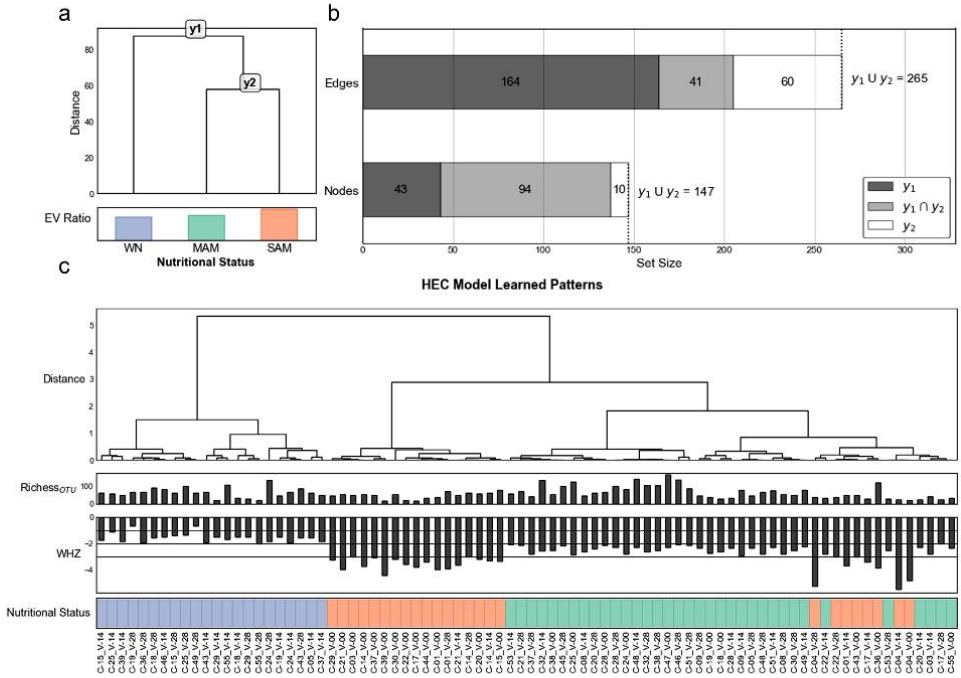


Figure 4. Hierarchical classification scheme for phenotypes using sample-specific perturbation networks, (a) Data-driven HEC model structure determined from perturbation matrix. (b) Barchart-styled Venn diagram showing node and edge sets selected by Clairvoyance. (c) Unsupervised clustering of prediction probabilities from HEC model on held-out LSOCV test sets.

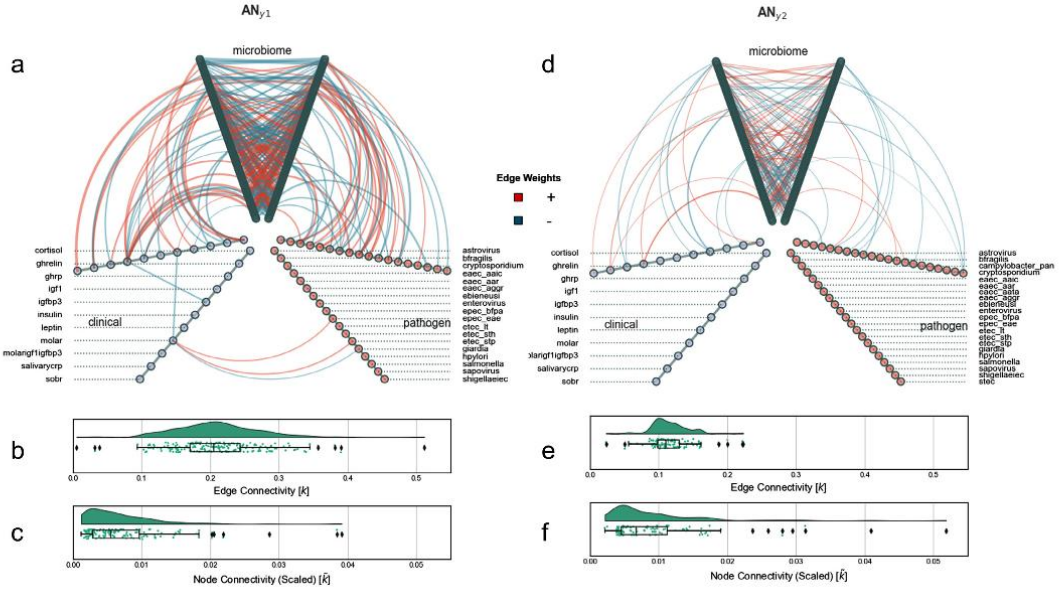


Figure 5. Hive plot network visualization of aggregate networks, Aggregate networks built using edges selected by Clairvoyance with edge weights from fitted logistic regression for sub-models y1 and y2 corresponding to AN_{y1} and AN_{y2}, respectively. Hive plots of (a) AN_{y1} and (d) AN_{y2} to visualize intra-modality and inter-modality connections where each axis represents a particular modality. Raincloud plots showing edge connectivity (b, e) and scaled node connectivity (c,f) of AN_{y1} and AN_{y2}, respectively.

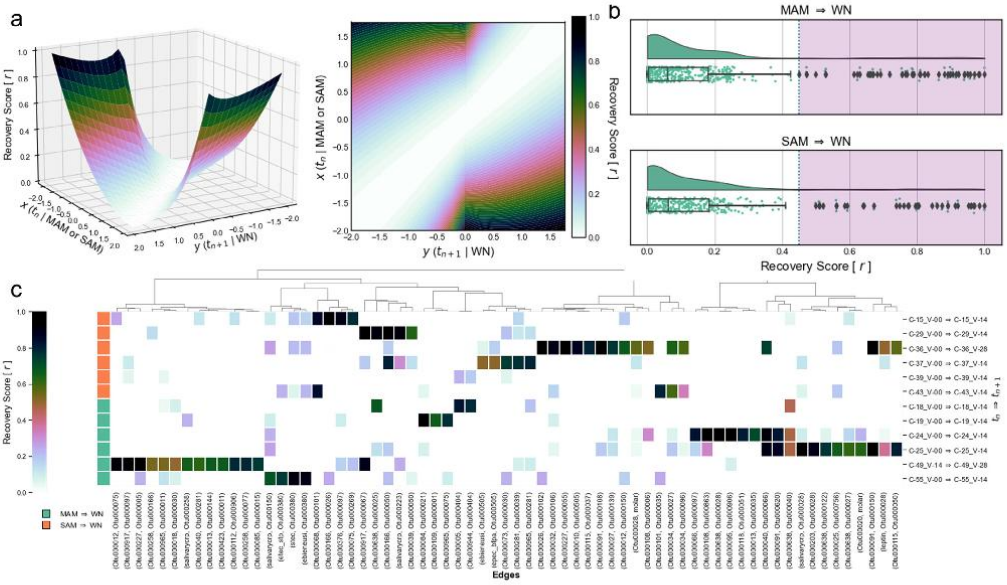


Figure 6. Recovery score for time-ordered SSPNs, (a) Domain and topology of recovery score in (left) 3-dimensions and (right) 2-dimensions. (b) Distributions of recovery scores for (top) MAM → WN and (bottom) SAM → WN transitions with boxplot whiskers representing IQR and violet overlay showcasing outlier regions (IQR*1.5 + Q3). (c) Clustermap showcasing recovery scores for outlier edges indicated in (b) and organized by transitional event.

immune development [12]. A new finding from our study was that children with MAM had statistically significant enrichments in the gut bacterial loads and variance in microbial richness compared to WN and SAM participants. These changes may be an indication of impaired immune function in the children with acute malnutrition [63] which agrees with our findings that IGF-1 and leptin are associated with various microbes and are highly predictive of nutritional status.

The aggregate network analyses suggest that WN and SAM are relatively more stable ecosystems and as a child transitions from WN to SAM they encounter a more chaotic and unpredictable microbiome. In addition to SAM having lower microbial diversity than MAM and WN, we also observed an enrichment in *Enterobacteriaceae* abundance therefore increasing the risk of adverse outcomes. Previous metagenomics studies have reported reduced microbial diversity and an increase in pathogenic *Enterobacteriaceae* among malnourished children [64].

Analysis of clinical measurements pertaining to energy regulating hormones gave insight into which metabolites were proportional to fluctuations in WHZ. We identified statistically significant trends between IGF-1, leptin, and IGFBP- 3 (the main binding protein of IGF-1) that increase with WHZ and nutritional status. IGF-1 is a key growth regulating hormone in infancy and plays an important role during nutritional recovery in undernourished children. Similarly, leptin is a hormone predominantly made by adipose cells and enterocytes in the small intestine to help regulate energy balance by moderating appetite and intestinal barrier function [65] and plays a major role in signaling energy deficit in acute malnutrition (Bouillanne et al., 2007; Prentice, Moore, Collinson, & O'Connell, 2002). We also observed reverse trends in that the levels of sOB-R (and the molar excess of sOB-R:leptin), ghrelin (and its binding protein) increase monotonically as WHZ decreases. Our findings support the hypothesis by Stein et al. 2006 that sOB-R is upregulated during starvation to maintain low levels of bioactive leptin and increase its half-life, thus, decreasing energy expenditure and increasing food uptake. Ghrelin is a well-studied hormone produced by enteroendocrine cells of the gastrointestinal, tract with substantial production in the stomach, [66,67] and circulating ghrelin blood levels are often highest when an individual experiences hunger while returning to lower levels after food intake [67,68]. Our clinical measurements of these metabolites are consistent with previous studies and provide a strong foundation for more complex analytical methods in the context of acute malnutrition.

Enterobacteriaceae abundance is greatly enriched in SAM children and may be linked to the low prevalence of *Giardia* which competes for the same ecological niche in the small intestine [69]. However, previous research using mouse models showed that Enterobacteriaceae is over-representation in Giardia infected mice [70]. There appears to be a more complex mechanism regulating the balance between Enterobacteriaceae (bacteria) and Giardia (protozoan parasite). Our finding may therefore be specific for marasmic SAM involving the differential regulation of anti-parasitic and anti-bacterial immune responses in these children where Giardia infection alters immune responses to *E.coli* or vice versa. This warrants further exploration. We also observed that EPEC virulence factors *bfpA* and *eae* increased with increasing severity of malnutrition. EPEC adheres to intestinal epithelial cells, causing diarrhea, and constitutes a significant risk to health, especially in very young children [71]. Subramanian and colleagues also reported an enrichment of *Enterobacteriaceae spp.* among children with SAM from Bangladesh [9,13]; although, a causal pathway is yet to be identified.

We found that *Escherichia/Shigella* sp. and molar ratio of sOB-R: letpin had substantially greater predictive capacity in discriminating WN from undernourished participant samples compared to other nodes. In particular, this *Escherichia-Shigella* OTU had high predictive capacity through its associations with *ghrelin* and *ghrp*. This is not surprising as leptin is a key player in regulating both antimicrobial peptides and microbiota composition and as such, *Escherichia-Shigella* and molar-excess soluble leptin may play pivotal roles in mediating complex interactions that modulate nutritional status. In the

TAC analyses, the gene target for *Escherichia/Shigella* was *ipaH* which is carried by *Shigella* sp. and EIEC. However, it should be noted that the signals from the *ipaH* gene target detected are most likely to have come from *Shigella* and not EIEC. Our previous studies showed that *Shigella flexneri* and *Shigella sonnei* account for the majority of ipaH detections[36].

We observed high predictive capacity of molar excess of sOB-R: leptin through *Lactobacillus mucosae,* an unclassified *Haemophilus* and an unclassified *Ruminococcaceae* UCG-002. Previous research has identified strong associations between leptin and *Lactobacillus* and it is believed that leptin can modulate gut microbiota by stimulating mucin production which may favor bacterial growth [72]. *Lactobacillus* has been shown to maintain intestinal homeostasis and is speculated to attenuate the pro-inflammatory signaling induced by *Shigella* after invasion of epithelial lining [73]. Similarly, previous research has ascertained that leptin supplementation resulted in a higher proportion of *Ruminococcaceae* [74]. To our knowledge, no research has investigated relationships between *Haemophilus* and leptin in the context of acute malnutrition. Another intriguing finding was the high predictive capacity of perturbations in *IGF-1* and an unclassified *Enterobacteriacea* in discriminating MAM from SAM. The high predictive capacity of perturbations in *IGF-1* and the *Enterobacteriaceae* associations are relevant as *Enterobacteriaceae* are often enriched in children who are wasted along with decreased plasma IGF-1 concentrations [75] and decreased concentrations of IGF-1 and IGFBP- 3 have been observed in underweight mice [76]. *Enterobacteriaceae* are often be enriched in undernourished individuals [64] and coupled with decreased concentrations of IGF-1 [75] and IGFBP- 3 [76]. These findings from other research groups agree with our results showing that IGF-1 and IGFBP- 3 concentrations decrease with WHZ and are lowest in SAM. As immunity is heavily impaired in children experiencing SAM [77], the predictive associations between *Enterobacteriacea* and IGF-1 are not surprising. However, it is not uncommon for children experiencing SAM to develop septicaemia [63,77]. Previous research has shown that patients with sepsis have low levels of IGF-1 inversely correlated with enteric bacterial load [78]. Hunninghake et al. 2010 also supposed that translocation of bacteria across the gastrointestinal tract may occur.

Our predictive functional profiling analysis found that phenylalanine metabolism and benzoate degradation pathways were enriched in SAM compared to WN. Several prior studies have investigated phenylalanine in the context of undernutrition in early childhood [79]. To our knowledge, this is the first indication that benzoate degradation may play a significant role in acute malnutrition but this needs further investigation. Other individual enzymes that appear to have been lost to some degree in the SAM condition that could influence host nutrition include sortase, tryptophanase, and butyrate kinase. While far more single enzymes are enriched in SAM, they are only sparingly enriched relative to those that have been lost. Finally, these functional results are merely predictions based on a computational tool [45], thus should be considered carefully without further validation using shotgun metagenomic analyses.

Acute malnutrition is a complex multifactorial disease with interplay between the gut microbiome, energy regulating hormones, and the presence of enteric pathogens. It appears that WN systems are stable but as a child's weight declines, approaching MAM, the community destabilizes with increased microbial diversity and interactions. As a child's nutritional status deteriorates the gut microbiota community becomes depleted and dominated by pathogenic *Enterobacteriaceae* in an ecological collapse as demonstrated by low bacterial load, and low microbial diversity. Using novel methods, we show there are potentially diagnostic interactions for each of these transitions. The methods introduced in this study build upon the existing SSN framework of Liu et al. 2016 to investigate patient-specific networks by extending an approach into a multimodal framework and by bootstrapping samples to obtain distributions of associations rather than single point values. This work not only provides an insight into dynamic multimodal systems in the context of acute malnutrition but also illuminates the potential avenues for diagnostics and therapeutics. The framework and methods introduced in this research can applied broadly across biological sciences.

Contributors

HMN, BAK, MA and AMP conceptualized and designed the study; HMN, MC coordinated the data collection; ATJ, RSB processed the stool samples; JL and ERH provided TAC array cards and interpretation, BAK, MA, JS, MB, RB, CO and AKS coordinated and undertook the enteric pathogen analyses and microbiome sequencing; JLE and CLD performed microbiome analysis, network analysis, software development, and statistical modelling. JLE, HMN, CLD and BAK verified the underlying data, drafted the initial manuscript; all authors reviewed, revised, and approved the final version of the manuscript.

We would like to thank the MRCG@LSHTM Nutrition Theme field and laboratory teams that supported the collection and processing of the specimens. Our thanks are also due to all the study participants and their families. We would also like to acknowledge Naisha Shah of *J. Craig Venter Institute* for her insight into multimodal sample-specific network analysis.

Data Sharing Statement

Materials (clinical and biological specimens) will be shared upon request for health research provided there are sufficient quantities, appropriate agreements and ethical approvals in place. Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contacts Brenda A. Kwambana-Adams (brenda.kwambana@ucl.ac.uk) and Christopher L. Dupont (cdupont@jcvi.org).

The 16S amplicon reads generated during this study are available at NCBI via SRA:SRR 14459253- 14459364 (BioSamples: SAMN19053066-19053178) under BioProject PRJNA727842. The code, methodologies, and tutorials for Ensemble Networks are open-sourced in our *EnsembleNetworkX* Python package (https://github.com/jolespin/ensemble_networkx) under the BSD-3 license. Metadata, datasets, and networks are available via https://doi.org/10.6084/m9.figshare.16733584.

Declaration of Competing Interest

The authors have no competing interests to declare.

Acknowledgments

The work was supported by the UK Medical Research Council (MRC; MC-A760-5QX00) and the UK Department for International Development (DFID) under the MRC/DFID Concordat agreement; Bill and Melinda Gates Foundation (OPP 1066932) and the National Institute of Medical Research (NIMR), UK. This network analysis was supported by NIH U54GH009824 [CLD] and NSF OCE-1558453 [CLD].

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.ebiom.2021.103644.

References

[1] Forbes GB. Joint FAO/WHO ad hoc Expert Committee, Energy and Protein Requirements, WHO Technical Report Series 522. Arch Pediatr Adolesc Med 1974;127:296.

[2] Black RE, Victora CG, Walker SP, Bhutta ZA, Christian P, De Onis M, et al. Maternal and child undernutrition and overweight in low-income and middle-income countries. Lancet 2013;382:427–51.

[3] World Health Organization. WHO child growth standards: length/height-for-age, weight-for-age, weight-for-length. weight-for-height and body mass index-for-age: methods and development 2006 https://www.who.int/publications/i/item/924154693X (accessed 6 May2021).

[4] Franco VHM, Hotta JKS, Jorge SM, Dos Santos JE. Plasma fatty acids in children with grade III protein-energy malnutrition in its different clinical forms: Marasmus, marasmic kwashiorkor, and kwashiorkor. J Trop Pediatr 1999;45:71–5.

[5] The Gambia Bureau of Statistics. The Gambia Multiple Indicator Cluster Survey 2018. Banjul 2019 https://mics-surveys-prod.s3.amazonaws.com/MICS6/West and Central Africa/Gambia/2018/Snapshots/The Gambia 2018 MICS Statistical Snapshots_English.pdf (accessed 29 Jul2020).

[6] National Nutrition Agency (NaNA)-Gambia, UNICEF GB of S. Gambia National Micronutrient Survey. Banjul 2018;2019 https://groundworkhealth.org/wp-content/uploads/2019/03/GNMS2018-Final-Report_190325.pdf (accessed 29 Jul2020).

[7] Velly H, Britton RA, Preidis GA. Mechanisms of cross-talk between the diet, the intestinal microbiome, and the undernourished host. Gut Microbes 2017;8:98–112.

[8] Nabwera HM, Fulford AJ, Moore SE, Prentice AM. Growth faltering in rural Gambian children after four decades of interventions: a retrospective cohort study. Lancet Glob Heal 2017;5:e208–16.

[9] Platts-Mills JA, Taniuchi M, Uddin MJ, Sobuz SU, Mahfuz M, Gaffar SMA, et al. Association between enteropathogens and malnutrition in children aged 6-23 mo in Bangladesh: A case-control study. Am J Clin Nutr 2017;105:1132–8.

[10] Richard SA, McCormick BJJ, Miller MA, Caulfield LE, Checkley W. MAL-ED Network Investigators. Modeling environmental influences on child growth in the MAL-ED cohort study: opportunities and challenges. Clin Infect Dis 2014;59(4):S255. Suppl.

[11] Keusch GT, Denno DM, Black RE, Duggan C, Guerrant RL, Lavery JV, et al. Environmental enteric dysfunction: Pathogenesis, diagnosis, and clinical consequences. Clin Infect Dis 2014;59:S207–12.

[12] Smith MI, Yatsunenko T, Manary MJ, Trehan I, Mkakosya R, Cheng J, et al. Gut microbiomes of Malawian twin pairs discordant for kwashiorkor. Science 2013;339(80):548–54.

[13] Subramanian S, Huq S, Yatsunenko T, Haque R, Mahfuz M, Alam MA, et al. Persistent gut microbiota immaturity in malnourished Bangladeshi children. Nature 2014;510:417–21.

[14] Kristensen KHS, Wiese M, Rytter MJH, Özçam M, Hansen LH, Namusoke H, et al. Gut Microbiota in Children Hospitalized with Oedematous and Non-Oedematous Severe Acute Malnutrition in Uganda. PLoS Negl Trop Dis 2016;10:e0004369.

[15] Gough EK, Stephens DA, Moodie EEM, Prendergast AJ, Stoltzfus RJ, Humphrey JH, et al. Linear growth faltering in infants is associated with Acidaminococcus sp. and community-level changes in the gut microbiota. Microbiome 2015;3. doi: 10.1186/s40168-015-0089-2.

[16] Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, et al. Human gut microbiome viewed across age and geography. Nature 2012;486:222–7.

[17] Yadav D, Ghosh TS, Mande SS. Global investigation of composition and interaction networks in gut microbiomes of individuals belonging to diverse geographies and age-groups. Gut Pathog 2016;8. doi: 10.1186/s13099-016-0099-z.

[18] Davis JCC, Lewis ZT, Krishnan S, Bernstein RM, Moore SE, Prentice AM, et al. Growth and Morbidity of Gambian Infants are Influenced by Maternal Milk Oligosaccharides and Infant Gut Microbiota. Sci Rep 2017;7:1–16.

[19] Kashyap PC, Chia N, Nelson H, Segal E, Elinav E. Microbiome at the Frontier of Personalized Medicine. Mayo Clin. Proc. 2017;92:1855–64.

[20] David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, et al. Diet rapidly and reproducibly alters the human gut microbiome. Nature 2014;505:559–63.

[21] Grzeskowiak L, Collado MC, Mangani C, Maleta K, Laitinen K, Ashorn P, et al. Distinct Gut microbiota in southeastern African and northern European infants. J Pediatr Gastroenterol Nutr 2012;54:812–6.

[22] De Filippo C, Cavalieri D, Di Paola M, Ramazzotti M, Poullet JB, Massart S, et al. Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. Proc Natl Acad Sci U S A 2010;107:14691–6.

[23] Monira S, Nakamura S, Gotoh K, Izutsu K, Watanabe H, Alam NH, et al. Gut microbiota of healthy and malnourished children in Bangladesh. Front Microbiol 2011;2. doi: 10.3389/fmicb.2011.00228.

[24] Blanton LV, Barratt MJ, Charbonneau MR, Ahmed T, Gordon JI. Childhood undernutrition, the gut microbiota, and microbiota-directed therapeutics. Science 2016;352(80):1533.

[25] Chen RY, Mostafa I, Hibberd MC, Das S, Mahfuz M, Naila NN, et al. A Microbiota-Directed Food Intervention for Undernourished Children. N Engl J Med 2021;384. doi: 10.1056/nejmoa2023294.

[26] Mostafa I, Nahar NN, Islam MM, Huq S, Mustafa M, Barratt M, et al. Proof-of-concept study of the efficacy of a microbiota-directed complementary food formulation (MDCF) for treating moderate acute malnutrition. BMC Public Health 2020;20. doi: 10.1186/s12889-020-8330-8.

[27] Gehrig JL, Venkatesh S, Chang HW, Hibberd MC, Kung VL, Cheng J, et al. Effects of microbiota-directed foods in gnotobiotic animals and undernourished children. Science 2019;365(80). doi: 10.1126/science.aau4732.

[28] Lunn PG, Erinoso HO, Northrop-Clewes CA, Boyce SA. Giardia intestinalis is unlikely to be a major cause of the poor growth of rural Gambian infants. J Nutr 1999;129:872.

[29] Campbell DI, McPhail G, Lunn PG, Elia M, Jeffries DJ. Intestinal inflammation measured by fecal neopterin in Gambian children with enteropathy: association with growth failure, Giardia lamblia, and intestinal permeability. J Pediatr Gastroenterol Nutr 2004;39:153.

[30] Thomas JE, Gibson GR, Darboe MK, Weaver LT, Dale A. Isolation of Helicobacter pylori from human faeces. Lancet 1992;340:1194–5.

[31] World Food Programme. WFP The Gambia - Country Brief 2020.

[32] Husseini M, Darboe MK, Moore SE, Nabwera HM, Prentice AM. Thresholds of socio-economic and environmental conditions necessary to escape from childhood malnutrition: A natural experiment in rural Gambia. BMC Med 2018;16. doi: 10.1186/s12916-018-1179-3.

[33] Nabwera HM, Bernstein RM, Agbla SC, Moore SE, Darboe MK, Colley M, et al. Hormonal Correlates and Predictors of Nutritional Recovery in Malnourished African Children. J Trop Pediatr 2018;64:364–72.

[34] World Health Organization. Pocket Book of Hospital Care for Children: Guidelines for the Management of Common Childhood Illnesses. World Health Organization 2013 http://www.who.int/maternal_child_adolescent/documents/9241546700/en/(accessed 5 May2021).

[35] Stein K, Vasquez-Garibay E, Kratzsch J, Romero-Velarde E, Jahreis G. Influence of Nutritional Recovery on the Leptin Axis in Severely Malnourished Children. J Clin Endocrinol Metab 2006;91:1021–6.

[36] Liu J, Gratz J, Amour C, Kibiki G, Becker S, Janaki L, et al. A laboratory-developed taqman array card for simultaneous detection of 19 enteropathogens. J Clin Microbiol 2013;51:472–80.

[37] Donnenberg MS, Girón JA, Nataro JP, Kaper JB. A plasmid-encoded type IV fimbrial gene of enteropathogenic Escherichia coli associated with localized adherence. Mol Microbiol 1992;6:3427–37.

[38] Edgar RC. UPARSE: Highly accurate OTU sequences from microbial amplicon reads. Nat Methods 2013;10:996–8.

[39] Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 2009;75:7537–41.

[40] Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS One 2010;5:e9490.

[41] Morton JT, Sanders J, Quinn RA, McDonald D, Gonzalez A, Vázquez-Baeza Y, et al. Balance Trees Reveal Microbial Niche Differentiation. mSystems 2017;2:e00162.

[42] Biocore. scikit-bio: A Bioinformatics Library for Data Scientists, Students, and Developers. GitHub 2020 https://github.com/biocore/scikit-bio (accessed 4 Jun2020).

[43] Espinoza JL. compositional: Compositional data analysis in Python. GitHub 2020 https://github.com/jolespin/compositional (accessed 15 May2020).

[44] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020;17:261–72.

[45] Douglas GM, Maffei VJ, Zaneveld JR, Yurgel SN, Brown JR, Taylor CM, et al. PICRUSt2 for prediction of metagenome functions. Nat. Biotechnol. 2020;38:685–8.

[46] Sun S, Jones RB, Fodor AA. Inference-based accuracy of metagenome prediction tools varies across sample types and functional categories. Microbiome 2020;8:46.

[47] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545–50.

[48] Fernandes AD, Macklaim JM, Linn TG, Reid G, Gloor GB. ANOVA-Like Differential Expression (ALDEx) Analysis for Mixed Population RNA-Seq. PLoS One 2013;8:e67019.

[49] Espinoza JL. soothsayer: High-level analysis package for (bio-)informatics. GitHub 2019 https://github.com/jolespin/soothsayer (accessed 7 Sep2019).

[50] Seabold S, Perktold J. Statsmodels: Econometric and Statistical Modeling with Python 2010 https://statsmodels.sourceforge.net/ (accessed 11 May2020).

[51] Espinoza JL, Shah N, Singh S, Nelson KE, Dupont CL. Applications of weighted association networks applied to compositional data in biology. Environ Microbiol 2020;22:3020–38.

[52] Erb I, Notredame C. How should we measure proportionality on relative gene expression data? Theory Biosci 2016;135:21–36.

[53] Lovell D, Pawlowsky-Glahn V, Egozcue JJ, Marguerat S, Bähler J. Proportionality: A Valid Alternative to Correlation for Relative Data. PLOS Comput Biol 2015;11:e1004075.

[54] Hagberg AA, Schult DA, Swart PJ. Exploring Network Structure, Dynamics, and Function using NetworkX 2008 http://math.lanl.gov/~hagberg/Papers/hagberg-2008-exploring.pdf (accessed 11 Jan2018).

[55] Espinoza JL. hive_networkx: Hive plots in Python. GitHub 2020 https://github.com/jolespin/hive_networkx (accessed 3 Aug2020).

[56] Krzywinski M, Birol I, Jones SJ, Marra MA. Hive plots–rational approach to visualizing networks. Brief Bioinform 2012;13:627–44.

[57] Allen M, Poggiali D, Whitaker K, Marshall TR, Kievit RA. Raincloud plots: A multi-platform tool for robust data visualization. Wellcome Open Res 2019;4. doi: 10.12688/wellcomeopenres.15191.1.

[58] Espinoza JL, Dupont CL, O'Rourke A, Beyhan S, Morales P, Spoering A, et al. Predicting antimicrobial mechanism-of-action from transcriptomes: A generalizable explainable artificial intelligence approach. PLOS Comput Biol 2021;17:e1008857.

[59] Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: A methodology review. J Biomed Inform 2002;35:352–9.

[60] Blank TE, Zhong H, Bell AL, Whittam TS, Donnenberg MS. Molecular variation among type IV pilin (bfpA) genes from diverse enteropathogenic Escherichia coli strains. Infect Immun 2000;68:7028–38.
[61] Slinger R, Lau K, Slinger M, Moldovan I, Chan F. Higher atypical enteropathogenic Escherichia coli (a-EPEC) bacterial loads in children with diarrhea are associated with PCR detection of the EHEC factor for adherence 1/lymphocyte inhibitory factor A (efa1/lifa) gene. Ann Clin Microbiol Antimicrob 2017;16:16.
[62] Alou MT, Million M, Traore SI, Mouelhi D, Khelaifia S, Bachar D, et al. Gut bacteria missing in severe acute malnutrition, can we identify potential probiotics by culturomics? Front Microbiol 2017;8. doi: 10.3389/fmicb.2017.00899.
[63] Jones KDJ, Berkley JA. Severe acute malnutrition and infection. Paediatr Int Child Health 2014;34:S1.
[64] Million M, Diallo A, Raoult D. Gut microbiota and malnutrition. Microb. Pathog. 2017;106:127–38.
[65] Brennan AM, Mantzoros CS. Drug Insight: The role of leptin in human physiology and pathophysiology - Emerging clinical applications. Nat. Clin. Pract. Endocrinol. Metab. 2006;2:318–27.
[66] Kojima M, Hosoda H, Date Y, Nakazato M, Matsuo H, Kangawa K. Ghrelin is a growth-hormone-releasing acylated peptide from stomach. Nature 1999;402:656–60.
[67] Müller TD, Nogueiras R, Andermann ML, Andrews ZB, Anker SD, Argente J, et al. Ghrelin. Mol. Metab. 2015;4:437–60.
[68] Cummings DE, Purnell JQ, Frayo RS, Schmidova K, Wisse BE, Weigle DS. A Preprandial Rise in Plasma Ghrelin Levels Suggests a Role in Meal Initiation in Humans. Diabetes 2001;50:1714–9.
[69] Allain T, Amat CB, Motta JP, Manko A, Buret AG. Interactions of Giardia sp. with the intestinal barrier: Epithelium, mucus, and microbiota. Tissue Barriers 2017;5. doi: 10.1080/21688370.2016.1274354.
[70] Bartelt LA, Bolick DT, Mayneris-Perxachs J, Kolling GL, Medlock GL, Zaenker EI, et al. Cross-modulation of pathogen-specific pathways enhances malnutrition during enteric co-infection with Giardia lamblia and enteroaggregative Escherichia coli. PLOS Pathog 2017;13:e1006471.
[71] Deborah Chen H, Frankel G. Enteropathogenic Escherichia coli: Unravelling pathogenesis. FEMS Microbiol. Rev. 2005;29:83–98.
[72] El Homsi M, Ducroc R, Claustre J, Jourdan G, Gertler A, Estienne M, et al. Leptin modulates the expression of secreted and membrane-associated mucins in colonic epithelial cells by targeting PKC, PI3K, and MAPK pathways. Am J Physiol - Gastrointest Liver Physiol 2007;293. doi: 10.1152/ajpgi.00091.2007.
[73] Tien M-T, Girardin SE, Regnault B, Le Bourhis L, Dillies M-A, Coppée J-Y, et al. Anti-Inflammatory Effect of Lactobacillus casei on Shigella -Infected Human Intestinal Epithelial Cells. J Immunol 2006;176:1228–37.
[74] Grases-Pintó B, Abril-Gil M, Castell M, Rodríguez-Lagunas MJ, Burleigh S, Fåk Hållenius F, et al. Influence of Leptin and Adiponectin Supplementation on Intraepithelial Lymphocyte and Microbiota Composition in Suckling Rats. Front Immunol 2019;10:2369.
[75] Bartz S, Mody A, Hornik C, Bain J, Muehlbauer M, Kiyimba T, et al. Severe acute malnutrition in childhood: Hormonal and metabolic status at presentation, response to treatment, and predictors of mortality. J Clin Endocrinol Metab 2014;99:2128–37.
[76] Schwarzer M, Makki K, Storelli G, Machuca-Gayet I, Srutkova D, Hermanova P, et al. Lactobacillus plantarum strain maintains growth of infant mice during chronic undernutrition. Science 2016;351(80):854–7.
[77] Hossain MI, Chisti J, Yoshimatsu S, Yasmin R, Ahmed T. Features in Septic Children With or Without Severe Acute Malnutrition and the Risk Factors of Mortality. Pediatrics 2015;135:S10.
[78] Hunninghake GW, Doerschug KC, Nymon AB, Schmidt GA, Meyerholz DK, Ashare A. Insulin-like growth factor-1 levels contribute to the development of bacterial translocation in sepsis. Am J Respir Crit Care Med 2010;182:518–25.
[79] Jahoor F, Badaloo A, Reid M, Forrester T. Sulfur amino acid metabolism in children with severe childhood undernutrition: Methionine kinetics. Am J Clin Nutr 2006;84:1400–5.
[80] Liu X, Wang Y, Ji H, Aihara K, Chen L. Personalized characterization of diseases using sample-specific networks. Nucleic Acids Res 2016;44:e164.
[81] Nataro JP, Martinez J. Diagnosis and Investigation of Diarrheagenic Escherichia coli. Methods Mol Med 1998;15:387–406.
[82] J H, AG T. Enteropathogenic Escherichia coli: foe or innocent bystander? Clin Microbiol Infect 2015;21:729–34.
[83] Asea A, Kaur P, Chakraborti A. Enteroaggregative Escherichia coli: An emerging enteric food borne pathogen. Interdiscip. Perspect. Infect. Dis. 2010;2010. doi: 10.1155/2010/254159.
[84] Fleckenstein JM, Kuhlmann FM. Enterotoxigenic Escherichia coli Infections. Curr. Infect. Dis. Rep. 2019;21:1–9.

Title:
Differential network analysis of oral microbiome metatranscriptomes identifies community scale metabolic restructuring in dental caries

Authors:
Josh L. Espinoza [1,2,3], Manolito G. Torralba[4], Pamela Leong[5], Richard Saffery[5], Michelle Bockmann[6], Claire Kuelbs[2], Suren Singh[3], Toby Hughes[6], Jeffrey M. Craig[5,7], Karen E. Nelson[2,4], Chris L. Dupont[1,2,*]

Affiliations:
1 Department of Environment and Sustainability, J. Craig Venter Institute, La Jolla, CA 92037, USA
2 Department of Human Biology and Genomic Medicine, J. Craig Venter Institute, La Jolla, CA 92037, USA
3 Applied Sciences, Durban University of Technology, Durban, South Africa
4 Department of Human Biology and Genomic Medicine, J. Craig Venter Institute, Rockville, MD 20850, USA
5 Epigenetics, Murdoch Children's Research Institute and Department of Paediatrics, The University of Melbourne, Parkville, 3052 Victoria Australia
6 Adelaide Dental School, The University of Adelaide
7 IMPACT Strategic Research Centre, Deakin University School of Medicine, Geelong, VIC 3220 Australia

Abstract:
Dental caries is a microbial disease and the most common chronic health condition, affecting nearly 3.5 billion people worldwide. In this study, we used a multi-omics approach to characterize the supragingival plaque microbiome of 91 Australian children, generating 658 bacterial and 189 viral metagenome assembled genomes with transcriptional profiling and gene-expression network analysis. We developed a reproducible pipeline for clustering sample-specific genomes to integrate metagenomics and metatranscriptomics analysis regardless of biosample overlap. We introduce novel feature engineering and compositionally-aware ensemble network frameworks while demonstrating their utility for investigating regime shifts associated with caries dysbiosis. These methods can be applied when differential abundance modeling does not capture statistical enrichments or the results from such analysis are not adequate for providing deeper insight into disease. We identified which organisms and metabolic pathways were central in a coexpression network as well as how these networks were rewired between caries and caries-free phenotypes. Our findings provide evidence of a core bacterial microbiome that was transcriptionally active in the supragingival plaque of all participants regardless of phenotype, but also show highly diagnostic changes in the ways that organisms interact. Specifically, many organisms exhibit high connectedness with central carbon metabolism to *Cardiobacterium* and this shift serves a bridge between phenotypes. Our evidence supports the hypothesis that caries is a multifactorial ecological disease.

Significance Statement:

Using metagenomics and metatranscriptomics, this study characterized 658 bacterial and 189 viral metagenome assembled genomes (MAG) from the supragingival plaque oral microbiome to investigate dental caries in 91 children. We developed methodologies for species-level clustering to characterize biologically accurate MAGs across non-overlapping samples. We also developed novel feature engineering and network analysis techniques which can be used to gain deeper insight into microbial diseases than differential abundance methods alone. With these new techniques, we identified regime shifts between caries and caries-free microbiomes where certain taxa switch their interactions with other organisms and metabolic pathways. Our study provides evidence for the hypothesis that caries is a multifactorial ecological disease and contains generalizable methods for microbiome research.

Correspondence:
Chris L. Dupont
cdupont@jcvi.org
Department of Environment and Sustainability, J. Craig Venter Institute, La Jolla, CA 92037, USA
Department of Human Biology and Genomic Medicine, J. Craig Venter Institute, La Jolla, CA 92037, USA

Introduction:

Oral diseases such as dental caries are a critical concern for public health. Untreated tooth decay is the most common chronic health condition and affects nearly 3.5 billion people worldwide (James et al., 2018). Tooth decay is most common in younger individuals with a prevalence of 20% of children aged 5 to 11 and 13% in adolescents aged 12 to 19, with low-income children being twice as likely to have cavities (Dye et al., 2012). In most low- and middle-income countries, the prevalence of oral diseases increases with urbanization and inadequate access to medical treatment (Peres et al., 2019). In high-income countries, dental treatment averages 5% of total health expenditure and 20% of out-of-pocket health expenditure (OECD, 2017) making the disease a socioeconomic issue for all. This silent epidemic has long been coupled with the rise of civilization as early evidence from the Pleistocene era suggests that agriculture and the exploitation of starchy plant foods have burdened mankind with carious lesions since early prehistory (Humphrey et al., 2014).

In the modern "ecological plaque hypothesis", oral diseases arise from environmental perturbations leading to a shift in the endogenous microbial community (Marsh, 1994) where the selection of pathogenic bacteria is coupled with the environment and any species with germane traits can contribute to pathogenesis (Takahashi & Nyvad, 2011). The evidence for this theory is largely from the advent of next generation sequencing (NGS) technologies and the ability to sequence uncultivated organisms. In the context of carious lesions, the environmental perturbation arises from persistent consumption of dietary sugars leading to a decrease in pH and, when sustained, shifts the population to a more aciduric and cariogenic microbial community which degrades the enamel (Kleinberg, 2002; Marsh & Bradshaw, 1997). Changing environments (e.g., a substantial increase in acidity) can destabilize previously stable microbial communities reconfiguring them into new stability domains, referred to as regime shifts (Folke et al., 2004)., such as a cariogenic microbiome (Nyvad & Takahashi, 2020). Furthermore, this regime shift of the oral microbiome towards a cariogenic state is the result of an environment partially created by the bacteria themselves creating a complex feed-forward loop. This complex feed-forward loop makes it difficult to diagnose the exact cause of each case and the development of therapeutics for severe cases.

Understanding the roles of microbes in the context of caries-related dysbiosis (from the perspective of human health and not microbial stability) is non-trivial and is often explored using association networks. Association networks such as coexpression (transcriptomics) or co-abundance (genomics) are powerful frameworks for investigating inferred biological interactions by grouping biological features, such as genes or microbes, with related metabolism (Smith et al., 2016) or complementary ecological niches (Gomez et al., 2017). Despite widescale usage, many approaches do not address NGS compositionality and this is a major concern because non-compositionally aware metrics (e.g., correlation) are known to yield spurious associations with no biological meaning (Lovell et al., 2015). Although packages such as WGCNA (Langfelder & Horvath, 2008) introduced intuitive and clever ways for analyzing fully-connected weighted gene association networks, they do not directly support compositionally-aware association metrics such as proportionality (Erb & Notredame, 2016; Lovell et al., 2015; Quinn et al., 2017); thus, the findings using such methods are based on a statistical fallacy. However, awareness of compositional data analysis (CoDA) has increasingly made its way from geology to bioinformatics (Erb & Notredame, 2016; Lovell et al., 2015; Morton et al., 2017; Quinn et al., 2018) with many advancements in the context of network analysis (Espinoza et al., 2020).

Association networks are often applied to individual organisms in controlled settings (Fuller et al., 2007) and extending these concepts to ecosystems introduces many challenges that arise from the complexity of the data where the exact abundances of biological features are often unknown *a priori* and the number of features increase by several orders of magnitude. Furthermore, as systems biology deals with interactions amongst biological features, the number of pairwise interactions scale quadratically. The vast number of variables in a microbiome not only makes hypothesis testing precarious but can also lead to statistical artifacts in downstream analysis due to the "curse-of-dimensionality" making interpretation exceedingly difficult (Altman & Krzywinski, 2018). Many dimensionality-reduction methods such as PCA, [N]MDS, t-SNE (Van Der Maaten, 2014), or UMAP (McInnes et al., 2018) lose accessibility to original biological features rendering interpretation limited and unintuitive. The genome-resolved hierarchical complexity of microbiomes results in dynamic distributions of expression or abundance influenced by other microbes and latent environmental variables not accounted for by the experimental design. These community-level

datasets require representations of the data that account for these abstractions and group genes within their genome-resolved structure; that is, explainable biological feature engineering.

The balance between biological accuracy and analytical practicality is a constant theme when pairing metagenomic assemblies with metatranscriptomic sequencing. On one hand, consensus assemblies and binning of metagenome assembles genomes (MAG) make it feasible to cross-reference genomic features across samples while providing a relatively dense counts table. Though not inherently sparse, sparsity in NGS technologies is common and a major hurdle in CoDA (Paulson et al., 2013; Silverman et al., 2020). This approach often produces user-friendly data structures, but the resulting MAGs will likely be composites of multiple *bona fide* microbial genomes of highly similar strains resulting in redundant and contaminated MAGs that lack true biological interpretation. On the other hand, assembling samples individually and binning genomes will produce more biologically accurate MAGs but mapping to these samples produces inherently sparse matrices that have an extremely large number of features, which is problematic for statistical analysis (Altman & Krzywinski, 2018), mostly filled with zeros because the *one-to-one* mapping of reads to highly redundant genomic features. When using paired metagenomics and metatranscriptomics to investigate dysbiotic systems, it is imperative to address these pitfalls by leveraging compositionally-aware network methodologies simultaneously with natural hierarchies inherent in the data.

Results:

Metagenome assembled genomes for bacteria and viruses

Metagenomes from 88 Australian children in this study were evaluated and analyzed previously (Espinoza et al., 2018; Shaiber & Eren, 2019) but substantial improvements in assembly, binning, and quality assessment methodologies warranted revisitation and reanalysis. After quality assessment, with our updated methods, we isolated 658 bacterial MAGs, 179 DNA viruses, and 10 RNA viruses (Fig. 1, Table S2). For clarity, DNA and RNA viruses refer to viruses derived from metagenomics or metatranscriptomic sequencing, respectively. These bacterial MAGs clustered (95% ANI where ANI refers to Average Nucleotide Identity) into 135 unique species-level clusters (SLC) representing 49 hitherto unclassified species with 26 of which classified as *Patescibacteria* candidate phyla radiation (CPR; 6 *Gracilibacteria/SR1*, 43 *Saccharibacteria*) and a total of 69 CPR MAGs collectively. Of the non-CPR SLCs, we identified 31 *Bacteroidota,* 22 *Proteobacteria*, 21 *Actinobacteriota*, 23 *Firmicutes,* 8 *Fusobacteriota,* and 4 *Campylobacterota* (Table 2,S2,S3). The DNA and RNA viruses clustered into 137 and 5 unique SLCs, respectively. Most of the DNA viruses were classified as *Caudovirales,* of unknown species, associated with the human oral (42 SLCs), gut (41 SLCs), human respiratory (1 SLC), and non-specific environments (1 SLC). Aside from these unknown species, we also identified several *Caudovirales* phages for *Arthrobacter* (7 SLCs), *Streptococcus* (4 SLCs), *Klebsiella*, *Haemophilus*, *Pasteurella*, *Pseudomonas*, and *Burkholderia*. Other than *Caudovirales*, we identified *Streptococcus* satellite phages (2 SLCs), unclassified CRESS-DNA *Parvovirus* associated with the human gut (2 SLCs), and an unclassified virus associated with the human oral environment. Most of the RNA viruses were *Escherichia* phages (4 SLCs) designated as Qbeta BZ1, MS2, and BZ13 strains but we also uncovered a novel virus with no close taxonomic classification. Only high-confidence viruses based on strict *CheckV* thresholds were considered for analysis to reduce false positives and increase interpretability. The bacterial and viral SLCs contained 64 and 113 singleton clusters; individual genomes that did not share 95% ANI with any other organisms in the dataset. No archaea, eukaryote, or novel bacterial genera beyond the CPR were detected.

Our metapantranscriptomics approach collapsed 1,248,783 ORFs into 255,737 SLC-specific orthogroups (247,943 bacterial and 7,794 viral) reducing the dimensionality by 80% with minimal loss in information content. By using SLC-specific orthogroups, we were able to maintain a “bag-of-genomes” paradigm, opposed to that of a “bag-of-genes” and preserving natural hierarchical structures inherent in ecology.

We observed only one taxonomic database discrepancy, *M-1507-144.A__MAXBIN2__bin.008* was classified by *GTDB-Tk* as a novel *Tannerella* but this MAG clustered in BC14 with 10 *Peptidiphaga sp000466175* with high confidence (> 95% ANI via *FastANI*) which suggests an update to the GCF_003033925.1 reference taxonomy in NCBI. This result is further strengthened by the *CheckM* basal classification of the *Actinobacteria* phylum.

Relative taxonomic expression and abundance

Clustering using genome-resolved gene expression grouped subjects into 5 distinct clusters (Fig. 2) but these expression patterns were not able to discriminate samples based on the presence or absence of caries. We also measured the silhouette scores using Aitchison distance against caries status and observed average scores close to zero ($|\bar{x}_{silhouette}| < 0.003$) for bacterial and viral microbiomes in both metatranscriptomics and metagenomics datasets indicating minimal phenotype partitioning capacity using individual features.

As these expression patterns are in center log-ratio (CLR), values close to 0 can be considered basal community-level expression and close to the geometric mean of expression values for the microbiome. We observed a core bacterial supragingival plaque microbiome at the genus level as almost every genus is transcriptionally active in every sample (Clusters 2.1-2.4, see *Methods* for naming scheme), regardless of phenotype, but this is not the case for either DNA or RNA viruses. Most of the viruses were detected with low expression and grouped in Cluster-2.5. In Cluster-2.5, there are a few DNA viruses that are detected in almost every sample including *Streptococcus satellite phage Jaavan335*, unclassified *Caudovirales* associated with human gut, *Burkholderia phage phiE255*, *Haemophilus phage SuMu,* and *Klebsiella phage ST405-OXA48phi1* with *Escherichia phage MS2* as the only high prevalence RNA virus. Cluster-2.4 contains an unclassified human oral DNA virus and an unclassified oral *Caudovirales* that are transcriptionally active in every sample at modest levels on par with many bacteria in the microbiome. Cluster-2.1, the cluster with highest overall transcript abundance, we observe mostly genera from *Bacteroidota*, *Proteobacteria*, *Firmicutes*(C), and *Fusobacteriota*. The most transcriptionally active genera in Cluster-2.1 are *Capnocytophaga, Streptococcus, Neisseria, Haemophilus_D, Aggregibacter, Porphyromonas, and Veillonella* with modest expression from other bacteria. Cluster-2.2, with CLRs of 1-2, includes genera that appear in downstream analysis including *Cardiobacterium*, *Corynebacterium*, and *Tannerella*. Cluster-2.3 is the cluster with baseline transcriptional activity (i.e., CLR close to 0) which contains all the *Saccharibacteria* and *SR1* CPR clade.

We also investigated the RNA:DNA ratios from the 26 overlapping metatranscriptomic and metagenomic samples (Fig. 3). Based on clustering of RNA:DNA ratios, 5 distinct groupings were observed (Clusters-3.1-5). The most prominent findings are from Clusters-3.1,2 where taxa have highest and lowest RNA:DNA ratios, respectively. Cluster-3.1, the most transcriptionally active, included *Haemophilus_A* and *Alloprevotella,* the most active genera, as well as *Aggregibacter, Gemella, and Campylobacter_A*. Cluster-3.2 has a more uniform distribution of low RNA:DNA ratios and contained *Gracilibacteria*, *Saccharibacteri*a, *Streptococcus satellite phage Javan335,* and an unclassified *Caudovirales* phage associated with the human gut. Clustering of these RNA:DNA ratios also did not differentiate subjects based on phenotype.

In terms of alpha diversity, we did not observe any difference in bacterial richness for metatranscriptomics ($\tilde{x}$ = 131 SLCs) or metagenomics ($\tilde{x}$ =134 SLCs) datasets between caries and caries-free microbiomes (Mann-Whitney p >> 0.05). For viral richness, we did not observe difference in the metatranscriptomics dataset ($\tilde{x}$ =26 SLC) but observed a slight enrichment in viral richness in the caries microbiome ($\tilde{x}_{Caries}$ = 34.5 SLCs, $\tilde{x}_{Caries\text{-}free}$ = 30 SLCs; Mann-Whitney p = 0.026).

We implemented differential expression analysis between caries and caries-free cohorts at the taxonomic level (SLC expression) and PGFC level (engineered taxonomy-functional composite features). We did not observe any statistically significant components, neither SLCs nor PGFCs, using compositionally-aware methods such as ANCOM and ALDEx2. However, the lack of clear taxonomic or functional differences between the cohorts suggests interactions between variables is important, illustrating the need of differential networks to interrogate the caries and caries-free microbial systems.

Phenotype-specific coexpression networks reveal unique taxonomic and metabolic characteristics

As the caries phenotype is a multifactorial disease (Takahashi & Nyvad, 2011), the most natural approach for investigating associations would be through network analysis as such methodologies are useful for modeling complex systems with unknown structure. To be specific, the true structure (if one exists) of microbial interactions within each phenotype in our dataset is unknown *a priori*. Therefore, we must infer the structure of each network using data-driven approaches. Using an ensemble approach, we computed compositionally-aware coexpression networks, with PGFCs, an engineered feature based on taxonomy and functional potential (see *Methods*), as nodes (N_{nodes} = 2,478) and *rho* proportionalities as edge weights

(N_{edges} = 3,069,003), for caries and caries-free microbiomes ($PSCN_{Caries}$ and $PSCN_{Caries\text{-}free}$, respectively). The total connectivity of the $PSCN_{Caries}$ was 301,163.9 *k* with $PSCN_{Caries\text{-}free}$ ~7% lower (279,832.1 *k,* Table S6). Unsupervised clustering of the PSCNs sorted by median connectivity revealed clusters heterogeneous with respect to taxonomy, and a sharp drop off in connectivity at 250 *k* (Fig. 4A,B, Table S5, Table S7). In this high connectivity range, there are 12 $PSCN_{Caries}$ clusters (749 PGFCs) and 8 $PSCN_{Caries\text{-}free}$ clusters (555 PGFCs) which will be referred to as high connectivity PSCN clusters (HCPC).

One approach in computing homogeneity is via normalized entropy and, in this context, can be interpreted as cluster homogeneity where low entropy translates to a cluster being dominated by a single taxa (more homogenous) and high entropy as taxa being evenly distributed within a cluster (more heterogeneous). The most highly connected cluster in both PSCNs is Cluster-1 (HCPC-4A.1 and HCPC-4B.1) which is the second largest cluster in each network and one of the most heterogenous with respect to taxonomy. We observed a modest trend that HCPCs in the caries microbiome have higher taxonomic homogeneity than the caries-free microbiome. The caries HCPCs tend to have lower normalized entropy than the caries-free HCPCs especially compared to when considering all clusters; though, the number of observations was not sufficient for this statistical analysis and these results will not be further explored.

Despite the caries microbiome HCPCs being slightly more homogenous, the highest connectivity $PSCN_{Caries}$ cluster (HCPC-4A.1) is one of the most heterogenous clusters in the system. The majority of HCPC-4A.1 connectivity (82%) is from *Veillonella*, *Streptococcus, Granulicatella,* and *Kingella_B.* The remaining caries HCPCs are enriched in other bacteria including *Streptococcus*, *Capnocytophaga*, *Haemophilus_D*, *Neisseria*, *Cardiobacterium*, and *Aggregatibacter*. The highest connectivity cluster in $PSCN_{Caries\text{-}free}$ is HCPC-4B.1 whose connectivity is primarily from *Streptococcus sanguinis*, *Veillonella parvula_A*, and *Granulicatella adiacens*. The remaining caries-free HCPCs are enriched in *Neisseria*, *Capnocytophaga*, *Fusobacterium*, *Haemophilus_D*, and *Prevotella*. We observed a substantial overlap in high connectivity genera but the cluster membership of these genera is phenotype-specific and these configurations may provide key insight into how a system, whether caries or caries-free, stabilizes.

The second highest connectivity HCPCs in both caries (HCPC-4A.32) and caries-free (HCPC-4B.22) microbiomes are homogenous for *Streptococcus* and *Neisseria*, respectively. Connectivity from HCPC-4A.32 is mainly derived from *S. sanguinis* (98%) while connectivity from HCPC-4B.22 is 100% attributable to a novel *Neisseria* species (BC6). We also observed HCPC-4A.39 as another homogenous caries HCPC for *Capnocytophaga sputigena.* In both PSCNs, carbohydrate metabolism (27% $PSCN_{Caries}$, 36% $PSCN_{Caries\text{-}free}$) and cofactor/vitamin biosynthesis (10.7% $PSCN_{Caries}$, 8.2% $PSCN_{Caries\text{-}free}$) are attributable to most of the connectivity (Fig. S2). Glycolysis, gluconeogenesis, and pentose phosphate are heterogenous amongst the HCPCs regardless of phenotype. The citric acid cycle was responsible for the majority of the carbohydrate connectivity in $PSCN_{Caries}$ HCPC-4A.29; a heterogenous cluster enriched in *Neisseria* and *Prevotella.* Several cofactor and vitamin metabolic pathways were common amongst the HCPCs.

Community detection algorithms such as Louvain (Blondel et al., 2008) and, its updated successor, Leiden (Traag et al., 2019) have been used to investigate the structure of large and complex networks. The former has been used to study various biological networks (Jackson et al., 2018; Wilson et al., 2017; XH & M, 2021; Zheng et al., 2021) while Leiden is new and sparingly applied to biological systems, it addresses defects associated with Louvain. As these algorithms are stochastic, we utilized an iterative version of the Leiden community detection algorithm to investigate how these phenotype-specific HCPCs are structured and how the HCPCs partition into tightly connected high-confidence communities in an induced graph. The caries HCPCs naturally partition into Communities-4C.I-III while the caries-free HCPCs partition into Communities-4D.I-II (Fig. 4C,D,Table S5). Leiden communities revealed similar coexpression of two complementary configurations in both PSCNs: 1) majority *Bacteroidota* ($PSCN_{Caries}$ Community-4C.I and $PSCN_{Caries\text{-}free}$ Community-4D.I); and 2) majority *Firmicutes* via *Streptococcus* ($PSCN_{Caries}$ Community-4C.II and $PSCN_{Caries\text{-}free}$ Community-4D.II) (Fig. 4E,F).

Community-4C.I and 4D.I have high overlap in taxonomic membership, but they also have several unique taxa that may provide insight into phenotype-specific system states. Interestingly, no *Neisseria* were observed in $PSCN_{Caries}$ Community-4C.I but high *Neisseria* genus-level membership was observed in the

complementary $PSCN_{Caries\text{-}free}$ Community-4D.I (Fig. 4E). However, in $PSCN_{Caries}$ we observed high *Neisseria* genus-level membership in Community-4C.III coexpressed with more *Neisseria* and *Haemophilus* (Fig. 4E). *Neisseria* are highly connected in both PSCNs but their community membership, the taxa they are interacting with, is phenotype specific. More specifically, *Neisseria* appears to shift from high coexpression with several *Bacteroidota* species in the caries-free cohort to other *Neisseria* and *Haemophilus* in the caries cohort. The connectivity of *Haemophilus_D parainfluenzae* and an unclassified *Neisseria* (BC6) is relatively high in $PSCN_{Caries}$ Community-4C.III and these taxa are completely absent from the *Neisseria* enriched community in $PSCN_{Caries\text{-}free}$.

In terms of metabolism, drug resistance is only observed in $PSCN_{Caries}$ Community 4C.II, specifically *Streptococcus sanguinis* beta-Lactam resistance. Both caries and caries-free microbiomes lack arginine, proline, and lipid metabolism in the *Bacteriodota*-centric communities ($PSCN_{Caries}$ Community-4C.I and $PSCN_{Caries\text{-}free}$ Community-4D.I) but provide these pathways in the *Firmicutes*-centric communities ($PSCN_{Caries}$ Community-4C.II and $PSCN_{Caries\text{-}free}$ Community-4D.II). Conversely, these *Firmicutes*-centric communities lack sulfur metabolism which appears to be provided by *Bacteriodota*-centric community. Central carbohydrate metabolism connectivity is much higher in the *Bacteriodota*-centric $PSCN_{Caries\text{-}free}$ Community-4D.I relative to all of the other communities which may suggest that the taxa and central carbohydrate mechanisms in this community promote a healthy oral microbiome.

We compared the scaled connectivity of different PGFC groupings between caries and caries-free PSCNs using the union of PGFCs in caries and caries-free HCPCs. We observed statistically significant differential connectivity when grouping PGFCs by taxonomic level (N = 7 PGFCs enriched in $PSCN_{Caries}$ and N = 11 PGFCS enriched in $PSCN_{Caries\text{-}free}$) and none when grouped by functional level (Fig. 5). Although, the connectivity of high-level metabolic functional profiles is similar for both PSCNs, the taxa responsible for these driving functions are unique to the phenotype. The taxa with enriched connectivity in $PSCN_{Caries}$ were *Kingella_B oralis* trailed by *Granulicatella adiacens*, *Haemophilus_D parainfluenzae*, *Capnocytophaga leadbetteri*, and *Streptococcus oralis.* The taxa with greatest enriched connectivity in $PSCN_{Caries\text{-}free}$ were *Streptococcus sanguinis*, *Abiotrophia sp001815873*, an unclassified *Neisseria* (BC6), and *Cardiobacterium hominis.* Although unclassified *Neisseria* and *Abiotrophia sp001815873* are enriched in $PSCN_{Caries\text{-}free}$, they are not present in the caries-free communities (Fig. 4E) because they were not in the caries-free HCPCs. This discrepancy in membership suggests that connectivities of these taxa, though enriched, were masked by other high connectivity taxa in $PSCN_{Caries\text{-}free}$.

Differential coexpression networks suggests community scale metabolic restructuring through Cardiobacterium hominis

Differential coexpression networks (DCN) reveal changes in connectivity between a reference (caries free) and treatment (caries) network. As ensemble PSCNs are the building blocks of DCNs, our DCNs provide the same benefits with respect to outlier resistance. Previous approaches have used DCNs but did not use compositionally-aware association metrics or ensemble networks (Fuller et al., 2007; Hsu et al., 2015). While differential abundance/expression analyses can be useful in identifying feature enrichment (e.g., OTU, MAG, ORF, gene, etc.), each method has their own caveats in assumptions about the data distributions (well characterized in (Morton et al., 2019) with the establishment of reference frames) and provide no information regarding differences in pairwise interactions; an essential perspective when studying diseases resulting from dysbiosis. Using the $PSCN_{Caries\text{-}free}$ as a reference network and $PSCN_{Caries}$ as the treatment network, we were able to construct a DCN using the 875 PGFCs from the community detection analysis for seamless cross-referencing between PSCNs and the DCN. In the DCN, differential connectivity (denoted as $\Delta\tilde{k}$) is positive and negative when a connectivity is enriched in the caries and caries-free microbiomes, respectively. Unsupervised clustering of the DCN revealed 6 clusters (Fig. 6, Table S5,6), of which there were 3 high connectivity DCN clusters (HCDC), each being diagnostic of phenotype; HCDC-6A.4 had enriched connectivity in the caries microbiome while HCDCs-6A.2 and 5 had enriched connectivity in the caries-free microbiome. For the only HCDC with connectivity enriched in the caries microbiome (HCDC-6A.4), the differential connectivity was primarily from *Capnocytophaga sputigena*, *Kingella_B oralis, Vellonella parvula_A*, *Streptococcus sanguinis*, *Streptococcus oralis*, and several species of unclassified *Streptococci* (Fig. S4A) via carbohydrate and cofactor/vitamin metabolism (Fig. S4B).

HCDC-6A.4 included 43% of all taxa within the DCN. HCDCs with enriched connectivity in caries-free microbiome contained a broader range of microbes. However, most of these taxa were in HCDC-6A.2 with more than 77% of the taxa in the DCN which was not the case for HCDC-6A.5 with 40% of the taxa. In HCDC-6A.2, most of the differential connectivity was attributable to *Streptococcus sanguinis, Abiotrophia sp001815873,* an unclassified *Neisseria* (BC6), *Rothia dentocariosa,* and several *Fusobacteriota* via carbohydrate metabolism, ATP synthesis, carbon fixation, and cofactor/vitamin biosynthesis (Fig. S4B). While HCDC-6A.2 is heterogenous in terms of taxa membership, HCDC-6A.5 is fairly homogenous with most of the connectivity from *Cardiobacterium hominis* via carbohydrate metabolism. *Veillonella parvula_A* and *Fusobacterium polymorphum* are the only taxa that had membership in all HCDCs. As *Veillonella parvula_A* had high differential connectivity in both caries and caries-free phenotypes through different metabolic pathways, this finding may provide insight into dysbiosis. Several other taxa including *Haemophilus_D parainfluenzae* had membership in the HCDC-6A.4 and with at least one HCDC with negative differential connectivity (Fig. S4A).

Comparing set membership between HCDCs revealed key metabolic differences between microbiomes. HCDC-6A.4 exclusively had 14 KEGG modules with the most notable including pentose phosphate pathway, phosphate acetyltransferase-acetate kinase, beta-Lactam resistance, several cofactor/vitamin pathways (Table S5). HCDC-6A.2 had 19 KEGG modules not in HCDC-6A.4 which included many carbohydrate metabolic, reductive pentose phosphate cycle, and dissimilatory nitrate reduction pathways. HCDC-6A.5 only had 4 exclusive KEGG modules including citrate cycle, fumarate reductase, and Raetz pathway with citrate cycle and fumarate reductase metabolism from *Cardiobacterium hominis*.

Hive plots are a network visualization framework that groups nodes with respect to predefined axes. In this case, grouping PGFCs by taxa or KEGG categories for higher-level nodes and HCDCs for axes. The hive structure visualizes both intra- and inter-cluster differential connectivity clearly revealing hub nodes connecting clusters (Fig. 6B,C). In the context of this DCN, *Cardiobacterium hominis* was a link between the highest differential connectivity HCDCs for caries (HCDC-6A.4) and caries-free (HCDC.6A.2) microbiomes even though each cluster's intra-cluster connectivity is sign specific. HCDC-6A.4 had very low connectivity to HCDC-6A.2 but both have high connectivity to HCDC-6A.5 primarily via *Cardiobacterium hominis*. However, positive differential connectivity from HCDC-6A.5 to HCDC-6A.4 was mainly from *Cardiobacterium hominis* carbohydrate metabolism and ATP synthesis from other bacterial species. In the connection between *Cardiobacterium hominis* and HCDC-6A.2, we observed many more taxa, also at greater differential connectivity magnitude, primarily through *Streptococcus sanguinis*, *Abiotrophia sp001815873*, and an unclassified *Neisseria* (BC6) with a long tail of taxa with negative differential connectivity. In this latter case, the highly negative differential connectivity from *Cardiobacterium hominis* to HCDC-6A.2 is spread out across many metabolic pathways and is not disproportionally weighted at carbohydrate and ATP synthesis suggesting *Cardiobacterium hominis* may have a holistic relationship in a caries-free microbiomes while also playing a potentially non-beneficial role in caries microbiomes. We expanded *Cardiobacterium hominis* carbohydrate metabolism and ATP synthesis modules out in a separate DCN (Fig. 7, Tables 3,S5,S7).

After removing low connectivity edges (Fig. S5), this DCN revealed 5 Leiden communities, denoted as Communities-7.I-V, with the 2 largest communities being Community-7.I and Community-7.II. Consistent with our previous hive networks, we observed a community with connectivity primarily enriched in the caries microbiome (Community-7.I) and several communities with connectivity almost exclusively enriched in caries-free microbiome (Communities-7.II-V). The exception to the latter is *Cardiobacterium hominis* pyruvate oxidation and *Capnocytophaga sputigena* polyketide sugar unit biosynthesis connectivity enriched in caries microbiome in Community-7.III. Community-7.II, the largest of the negative differential connectivity communities, was far less complex than Community-7.I and has *Cardiobacterium hominis* pentose phosphate as highly central nodes. There were 3 other small communities with negative differential connectivity and the most interesting of these is Community-7.IV as *Cardiobacterium hominis* glucose to UDP-glucose conversion is connected to mostly to *Veillonella parvula_A* cofactor/vitamin metabolism but also *Neisseria* nitrogen metabolism/methane metabolism and *Corynebacterium durum* carbohydrate metabolism.

The most complex and informative community is Community-7.I, which is primarily composed of positive differential connectivity edges, those enriched in the caries microbiome. The negative differential connectivity edges are primarily from ATP synthesis of *Veillonella parvula_A* and *Leptotrichia_A sp001274535*. Said nodes only have negative differential connectivity edges which suggest they are influential to the rest of the community in a caries-free microbiome and this may provide insight into community-scale restructuring in the caries microbiome. Also worthy of note, the only nodes with both positive and negative differential connectivity edges are from *Cardiobacterium hominis* supporting the hypothesis that *Cardiobacterium hominis* is an essential player in the transition from caries-free to caries phenotypes and vice versa. However, the most striking feature of this community is that *Cardiobacterium hominis* citrate cycle and fumarate reductase are highly centralized suggesting a shift in carbohydrate metabolism from pentose phosphate cycle to citrate acid cycle in the caries microbiome. We also observed various types of carbohydrate metabolism in Community-7.I with positive differential connectivity from several other organisms (Table 3).

Predictive models applied to caries diagnosis

Feature selection and predictive modeling was implemented to further evaluate PGFC features that were indicative of caries diagnosis. In particular, the *Clairvoyance* feature selection algorithm (Espinoza et al., 2021) that has been previously evaluated on identifying diagnostic genes related to antibiotic resistance (Espinoza et al., 2021) and multimodal associations related to childhood undernutrition (Nabwera et al., 2021) was used to identify PGFCs that were able to accurately discriminate caries individuals from caries-free individuals. To allow for seamless interpretation with the network analysis, the set of 212 PGFCs from the DCN were used as input into the feature selection algorithm and this was implemented for PGFCs represented as MCR and as CLR transformed abundances to yield two separate feature sets. This mixed feature architecture allowed for a novel type stacking ensemble where each base classifier uses a specific feature set and feature representation (e.g., MCR and CLR values simultaneously) leveraging the strength of each measurement in the ability to predict caries phenotype. The MCR feature set included 36 PGFCs, the CLR feature set included 27 PGFCs, and 11 PGFCs were shared between both base models (52 unique PGFCs) (Fig. S6B). The PGFCs selected via feature selection included *Cardibacterium hominis* pentose phosphate and TCA cycle as some of the highest weighted features that were able to discriminate caries phenotypes (Table S9). The stacking ensemble classification model was able to predict unobserved twin groupings with an accuracy of >96.5% (see *Methods* for cross-validation). In this context, accuracy can be interpreted as the reliability of a feature set to be sufficient in diagnosing caries. This is in stark contrast to predictive modeling using the 212 PGFCs from the DCN without feature selection which yielded a baseline classification accuracy of only 58.8%.

Discussion:

This study provides evidence of a core bacterial microbiome and a personalized viral microbiome that is transcriptionally active within the supragingival plaque of this cohort of Australian children regardless of collection center, age, or sex. This core microbiome supports the ecological plaque hypothesis that environmental conditions influence the metabolism of existing microbes nudging the community into a cariogenic configuration, rather than it being associated with extensive gain or loss of taxa. As the oral community is able to shift the collective metabolism to adapt to a cariogenic environment, the reverse must also be true given the prevalence of this core community. The implications of such a finding are both therapeutic and diagnostic. The specific abundances of taxa or even transcripts are not diagnostic, but a panel of transcripts and their associations are highly diagnostic and could be used as remote dentistry as demonstrated by the predictive model's high accuracy in diagnosing caries phenotype. Similarly, probiotics that revert community associations might be powerful therapeutics. Characterizing the interactions between microbes and their additive metabolism is expected to provide a deeper insight into what it means metabolically to have a cariogenic oral environment and, also important, a caries-free environment. One objective of the study was to determine if an untargeted tooth swab of both cariogenic and non-cariogenic communities combined with sequencing and *in silico* analysis could predict the signals diagnostic of phenotype with an accuracy >96.% using 52 unique biological features. This type of "take-at-home" assay augments current dental prophylaxis, which is dependent upon in-person visitation, and increases patient equity.

The complexity of this study required many novel methods to be developed for identifying mechanisms involved in caries-related dysbiosis. With a scope only considering single taxa expression patterns (e.g., unsupervised clustering of samples and differential expression analysis), we would have not been able to identify any distinguishing features of caries or caries-free phenotypes. Our development of novel association network methodologies built on the fundamental network concepts inspired by WGCNA such as implementing fully-connected, undirected, and weighted networks that can be clustered via hierarchical clustering. Our approach augments legacy methods by leveraging feature engineering to reduce dimensionality and exploitation of natural biological ontologies, the use of proportionality instead of correlation, consensus Leiden community detection, and a novel ensemble network framework to build distributions of edge weights rather than singular point estimates. Only through the inferred interactions between PGFCs were we able to notice trends in the data that could describe caries-related regime shifts. We averted the intractability of naive "bag-of-genes" representations of the data (i.e., ORFs as individual units) by flexible feature engineering methodologies to simulate an *in silico* reconstruction of the microbial community grouped by taxonomy and function; a "bag-of-genomes" approach. Analyzing caries-related dysbiosis using differential networks in PGFC-space rather than gene-space allowed for rapid and computationally economical methods for packing and unpacking biological hierarchies to explore regime shifts; bridging the gap between machine intelligence and biological insight. Although the feature engineering methods in this study are largely dependent on curated metabolic pathways, they were designed to be generalizable to custom databases and unsupervised paradigms such as hierarchical clustering paired with gene ontology analysis.

Exploring the oral microbiome from unique vantage points through analyzing networks specific to a phenotype, the comparison of connectivity profiles, and the differentials between networks provided insight into not only dysbiotic regime shifts but also maintenance of caries-free status. Even high-level network statistics are biologically relevant in the context of caries-related dysbiosis. For instance, the caries PSCN had substantially greater total connectivity than the caries-free PSCN, which can be interpreted as a higher number of interacting microbes and diverse metabolic pathways. This enrichment in inferred interactions within a diseased community relative to a non-diseased community has been observed in other forms of dysbiosis in the human gut microbiome such as inflammatory bowel disease and obesity (Chen et al., 2020). This finding is especially relevant considering the microbial richness does not differ between caries and caries-free microbiomes. The larger total connectivity and number of high connectivity coexpression clusters in caries microbiome suggests that there are more microbial and metabolic interactions occurring in carious systems. Similarly, the fewer number of high connectivity clusters in a caries-free microbiome suggests that the caries-free phenotype has a microbiome dominated by a few key taxa and metabolic pathways.

Neisseria appears to be a key player with high connectivity in the supragingival plaque oral microbiome regardless of caries phenotype. Previous research has observed *Neisseria* as highly abundant in both caries and caries-free microbiomes (Yang et al., 2021) but, to our knowledge, this study is the first to report this trend in the context of network connectivity (Fig. 4, 6) and RNA:DNA (Fig. 3). Although the connectivity of *Neisseria* is comparable in both microbiomes, the high connectivity in the caries microbiome is masked by a plethora of other highly connected genera and is ranked higher in the caries-free microbiome as a result of fewer high connectivity genera (Fig. 4A,B). However, we observed different microbial communities interacting with *Neisseria* when comparing between caries and caries-free microbiomes. In particular, several species of *Neisseria* were interacting with members of *Bacteroidota* in the caries-free microbiome and shifts to interactions with *Haemophilus_D parainfluenzae* and fellow *Neisseria* (mostly an unclassified *Neisseria* (BC6) and *Neisseria mucosa*) in the caries microbiome (Fig. 4E). This is interesting because several species of *Neisseria* had enriched connectivity in the caries-free microbiome and *Haemophilus_D parainfluenzae* had enriched connectivity in the caries microbiome (Fig. 5A,C). Although, *Neisseria* and *Haemophilus parainfluenzae* are both common in the oral cavity of caries-free individuals from the perspective of abundance (BJ et al., 2008; Liljemark et al., 1984; Zaura et al., 2009), their interactions with other coexpressed microbes known to be associated with infections in humans (e.g. *Prevotella conceptionensis* from Community-C.II (Amat et al., 2020; N et al., 2018)*,* may be indicative of caries dysbiosis. Many of the organisms discussed in this research have not been exhaustively characterized in the context of dental caries from an ecological perspective which presents an opportunity for future co-culture experiments.

The ability to collapse and expand PGFCs in these abstract network spaces can be used to identify unanticipated players with uncharacterized interactions relevant to maintaining either caries-free or caries microbiomes. For instance, when comparing PSCNs *Cardiobacterium hominis* is revealed to be one of the microbes with the highest enrichment in connectivity in the caries-free microbiome (Fig. 5A,C). However, the narrative is more complex when partitioning the PGFCs by differentially connected clusters (Fig. 6) and collapsing PGFCs by taxa-specific higher order KEGG categories. In a hive network layout, *Cardiobacterium hominis* emerges as a hub not only in the caries-free microbiome but also in the caries microbiome as it constitutes the majority of the differential connectivity within HCDC.5 primarily through ATP synthesis and carbohydrate metabolism. With *Cardiobacterium hominis* ATP synthesis and carbohydrate metabolism as a focal point, we were able to expand our focus to more specific KEGG modules while retaining high-level KEGG categories for the other microbes in the network to avoid the infamous and uninformative hairball plots (Krzywinski et al., 2012) of overly complex networks (Fig. 7).

This hierarchical network, further validated through predictive modeling, implicates *Cardiobacterium hominis* as a nexus between caries-free and caries dysbiotic states through a switch from pentose phosphate to TCA cycle carbohydrate metabolism. Previous metabolic research confirms that both the TCA cycle and the pentose phosphate pathway function within the supragingival plaque *in vivo* and glycolytic activation caused an increase in pentose phosphate activity (Takahashi et al., 2010). These findings suggest that *Cardiobacterium hominis* mediated pentose phosphate pathway metabolism promotes a caries-free microbiome with the support of *Streptococcus sanguinis* lysine metabolism, *Abiotrophia sp001815873* ATP synthesis, and *Neisseria* cofactor metabolism (Community-7.II). This hypothesis agrees with previous research as *Streptococcus sanguinis* and *Abiotrophia* have been known to cooccur in caries-free children (Kanasi et al., 2010) while *Neisseria,* as mentioned previously, has been associated with beneficial oral health. The simplicity of interactions enriched in the caries-free microbiome, Communities-7.II-V, agree with our theory that fewer taxa with more defined metabolisms are indicative of stable and healthy oral communities; thus, opening the door for potential probiotics, engineered microbial communities, and therapeutics for oral health and resilience.

The evidence for *Cardiobacterium hominis* TCA cycle and its association with caries dysbiosis is more complex in Community-7.I which has considerably more taxa and metabolic pathways than communities that include the pentose phosphate pathway. However, this agrees with our earlier finding that caries-related regime shifts include more high connectivity interactions without an increase in microbial richness; that is, greater total connectivity with the same core microbiome. Previous research has shown that the caries microbiome has the potential to metabolize more diverse sugar sources than the caries-free microbiome (Espinoza et al., 2018), which supports the notion that metabolism associated with dysbiotic caries communities is more complex than healthy communities without dental caries and, therefore, higher total network connectivity. In Communities-7.I-V, *Cardiobacterium hominis* is the only microbe that has connectivities enriched in caries and in caries-free microbiomes which supports our hypothesis of turncoat behavior in regards to oral health. *Cardiobacterium hominis* TCA cycle had enriched connectivity to carbohydrate metabolism from *Kingella oralis* (Cherkasov et al., 2019), *Streptococcus oralis* (Kanasi et al., 2010), and *Corynebacterium matruchotii* (Yang et al., 2021) in the caries microbiome which have previously been statistically associated with caries dysbiosis in children.

Diseases stemming from microbial dysbiosis are often complex and difficult to investigate due to computational limitations and human interpretation. Our research addresses a critical limitation in paired metagenomics and metatranscriptomics across multiple samples/subjects: that is, how to have biologically accurate assemblies not biased by coassembled chimeras while also producing overlapping features (e.g., SLC, SLC-specific orthogroups). Furthermore, the analytical methodology (e.g., feature engineering and network analysis) employed in this study, though developed for the oral microbiome, can be generalized to other diseases, environments, and modalities (e.g., clinical measurements, metabolomics, proteomics). This research demonstrates how investigating microbiomes from different vantage points can provide insight into microbial ecosystems and their relevance in health and disease.

Methods:

Study design and patient cohort overview used for metatranscriptomics

The primary objective of this study was to characterize the supragingival dental plaque of children with and without caries. A secondary objective assess how host genetics affects the microbiome in relation to caries and was not prioritized in this study due to the environmental effects on transcription. This study has been analyzed using 16S rRNA amplicon (Freire et al., 2020; Gomez et al., 2017) and shotgun metagenomics (Espinoza et al., 2018; Shaiber & Eren, 2019) datasets. The metatranscriptomics dataset in this publication has not been analyzed prior to this study. Our metagenomics dataset contains 88 subjects and our metatranscriptomics dataset contains 91 subjects with an overlap of 26 subjects. For each subject, there is a single sample per omics technology and 4 were removed by our strict quality control assessment for a total of 87 metatranscriptomic samples.

The study design has been described previously for this BioProject (PRJNA383868) in sister studies (Espinoza et al., 2018; Freire et al., 2020; Gomez et al., 2017). In particular for the metatranscriptomics cohort, dental plaque samples were collected from participants of the University of Adelaide Craniofacial Biology Research Group Tooth Emergence and Oral Health Study (CBRG) (n = 52), and the Murdoch Children's Research Institute (MCRI) Peri/Postnatal Epigenetic Twins Study (PETS) (n = 39). Human research with PETS subjects was approved by the Royal Children's Hospital Human Research Ethics Committee (#3174), and the CBRG cohort was approved by The University of Adelaide Human Research Ethics Committee (#H-2013-097). Research at the J. Craig Venter Institute was approved by the JCVI Institutional Review Board (#2013-182). All research was performed according to the listed institutions guidelines and informed consent was obtained from all participants' parent and/or legal guardians. Inclusion criteria included 5-11-year-old twins whose parents consented to this portion of the study.

Supragingival plaque samples were obtained at the commencement of a dental examination. Prior to sample collection, participants were guided not to brush their teeth the night preceding the sample collection nor on the day of sample collection. Metadata were collected from three separate questionnaires completed by the parents during the period from consent to prior to the dental examination being undertaken. The clinical questionnaires consisted of a total of 132 questions to survey oral and medical health, dietary patterns, and development patterns, and dental hygiene. Visual inspection of the oral cavity followed International Caries Detection and Assessment System (ICDAS II). The ICDAS II was used to assess and define dental caries at the initial and early enamel lesion stages through to dentin and more advanced stages of the disease. Examiners were experienced clinicians who had undergone rigorous calibration and were routinely recalibrated across measurement sites to minimize error. Caries history in each participant was initially reduced to a whole-mouth score and three classifications were utilized: no evidence of current or previous caries experience; evidence of current caries affecting the enamel layer only on one or more tooth surfaces; evidence of previous or current caries experience that has progressed through the enamel layer to involve the dentin on one or more tooth surfaces (including restorations or tooth extractions due to caries). For the purpose of phenotypic analysis, we classified disease states from twins as evidence of caries in enamel or dentin.

Our protocol samples the supragingival plaque of all teeth in the oral cavity during sample collection regardless of whether or not a tooth is suspected of containing a cavity. Although this yields a mixture of caries and caries-free communities it provides a powerful opportunity to develop diagnostic "at-home" tests where samples would be collected by patients. Our quality-controlled cohort consists of 36 caries and 51 caries-free samples sampled in this method.

Bioinformatics and data analysis

Please refer to *Supplemental Methods* for detailed descriptions on computational and analytical methodologies implemented in this study.

Tables:

Metadata	Caries	Caries-free

Age(μ, σ^2, min, max)	(8.09, 2.7, 5.5, 10.9)	(7.82, 2.21, 5.4, 10.8)
Sex(Female)	19	31
Sex(Male)	17	20
Center(CBRG)	21	28
Center(MCRI)	15	23
Total Samples	36	51

Table 1 – Metatranscriptomics cohort sample metadata
Overview of sample size for cohort with respect to phenotype and several metadata.

Bacterial Phyla	SLCs	MAGs	Novel Species (SLCs)
p__Actinobacteriota	21	153	3
p__Bacteroidota	31	160	13
p__Campylobacterota	4	4	3
p__Firmicutes	16	35	4
p__Firmicutes_A	4	6	2
p__Firmicutes_C	3	62	0
p__Fusobacteriota	8	40	1
p__Patescibacteria	26	69	18
p__Proteobacteria	22	129	5
Total	135	658	49

Table 2 – Bacterial SLCs and MAGs with respect to phylum
The number of bacterial MAGs, SLCs, and novel species with respect to phyla.

PGFC	Species	Category	Description	HCDC	Community
BC123|M00150	*Cardiobacterium hominis*	ATP	Fumarate reductase, prokaryotes	5	I
BC46|M00157	*Granulicatella adiacens*	ATP	F-type ATPase, prokaryotes and chloroplasts	4	I
BC60|M00144	*Kingella_B oralis*	ATP	NADH:quinone oxidoreductase, prokaryotes	4	I
BC21|M00159	*Leptotrichia_A sp001274535*	ATP	V-type ATPase, prokaryotes	2	I
BC1|M00153	*Veillonella parvula_A*	ATP	Cytochrome bd ubiquinol oxidase	2	I
BC123|M00009	*Cardiobacterium hominis*	CCM	Citrate cycle (TCA cycle, Krebs cycle)	5	I
BC123|M00011	*Cardiobacterium hominis*	CCM	Citrate cycle, second carbon oxidation, 2-oxoglutarate => oxaloacetate	5	I
BC0b|M00003	*Corynebacterium matruchotii*	CCM	Gluconeogenesis, oxaloacetate => fructose-6P	4	I
BC0b|M00001	*Corynebacterium matruchotii*	CCM	Glycolysis (Embden-Meyerhof pathway), glucose => pyruvate	4	I
BC0b|M00002	*Corynebacterium matruchotii*	CCM	Glycolysis, core module involving three-carbon compounds	4	I
BC46|M00007	*Granulicatella adiacens*	CCM	Pentose phosphate pathway, non-oxidative phase, fructose 6P => ribose 5P	4	I
BC60|M00011	*Kingella_B oralis*	CCM	Citrate cycle, second carbon oxidation, 2-oxoglutarate => oxaloacetate	4	I
BC60|M00003	*Kingella_B oralis*	CCM	Gluconeogenesis, oxaloacetate => fructose-6P	4	I
BC60|M00001	*Kingella_B oralis*	CCM	Glycolysis (Embden-Meyerhof pathway), glucose => pyruvate	4	I
BC60|M00002	*Kingella_B oralis*	CCM	Glycolysis, core module involving three-carbon compounds	4	I
BC60|M00307	*Kingella_B oralis*	CCM	Pyruvate oxidation, pyruvate => acetyl-CoA	4	I
BC60|M00308	*Kingella_B oralis*	CCM	Semi-phosphorylative Entner-Doudoroff pathway, gluconate => glycerate-3P	4	I
BC36|M00005	*Streptococcus oralis*	CCM	PRPP biosynthesis, ribose 5P => PRPP	4	I
BC46|M00854	*Granulicatella adiacens*	OCM	Glycogen biosynthesis, glucose-1P => glycogen/starch	4	I
BC18|M00157	*Abiotrophia sp001815873*	ATP	F-type ATPase, prokaryotes and chloroplasts	2	II

BC18|M00159	*Abiotrophia sp001815873*	ATP	V-type ATPase, prokaryotes	2	II
BC123|M00004	*Cardiobacterium hominis*	CCM	Pentose phosphate pathway (Pentose phosphate cycle)	5	II
BC123|M00007	*Cardiobacterium hominis*	CCM	Pentose phosphate pathway, non-oxidative phase, fructose 6P => ribose 5P	5	II
BC1|M00003	*Veillonella parvula_A*	CCM	Gluconeogenesis, oxaloacetate => fructose-6P	5	II
BC1|M00001	*Veillonella parvula_A*	CCM	Glycolysis (Embden-Meyerhof pathway), glucose => pyruvate	5	II
BC1|M00002	*Veillonella parvula_A*	CCM	Glycolysis, core module involving three-carbon compounds	5	II
BC123|M00854	*Cardiobacterium hominis*	OCM	Glycogen biosynthesis, glucose-1P => glycogen/starch	5	III
BC123|M00307	*Cardiobacterium hominis*	CCM	Pyruvate oxidation, pyruvate => acetyl-CoA	5	IV
BC123|M00549	*Cardiobacterium hominis*	OCM	Nucleotide sugar biosynthesis, glucose => UDP-glucose	5	V

Table 3 – Carbohydrate metabolism and ATP synthesis nodes in DCN Leiden Communities
Category refers to KEGG Level 3 metabolic category while description refers to KEGG module description. Community refers to Leiden communities for DCN.

Acronyms: ATP – ATP Synthesis, CCM – Central carbohydrate metabolism, OCM – Other carbohydrate metabolism.

Figures:

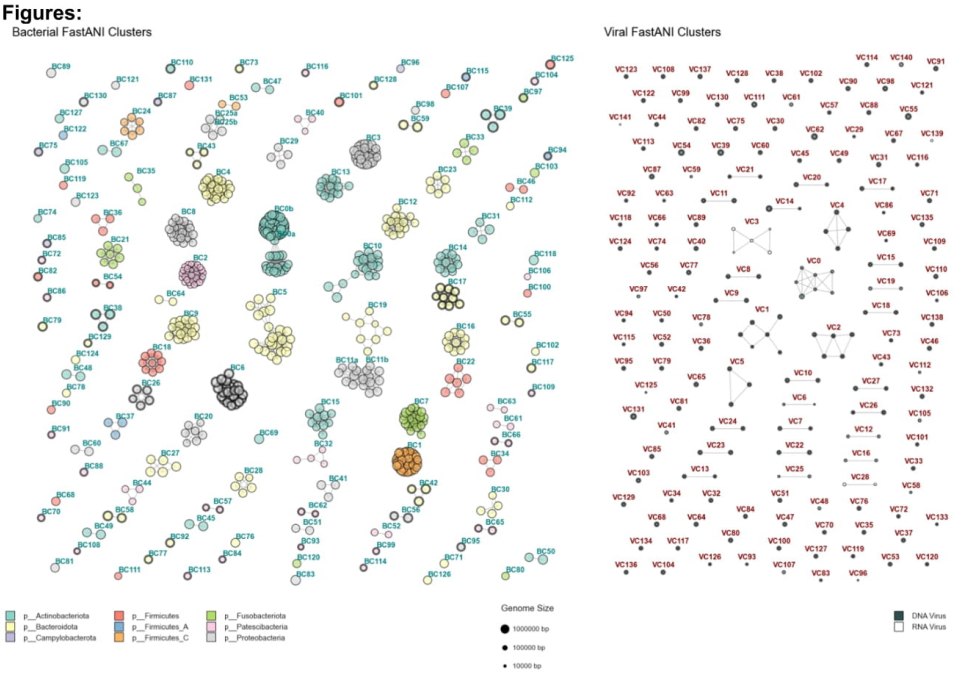


Fig. 1 – Network of *FastANI* clusters for bacterial and viral MAGs
Network with edge weights corresponding to ANI and nodes representing MAGs. (Left) Bacterial MAGs colored by phylum. (Right) Viral MAGs colored by either DNA or RNA virus type. Thicker edge weights indicate novel species not found in *GTDB-Tk* or *CheckV* for bacterial and viral *FastANI* clusters, respectively.

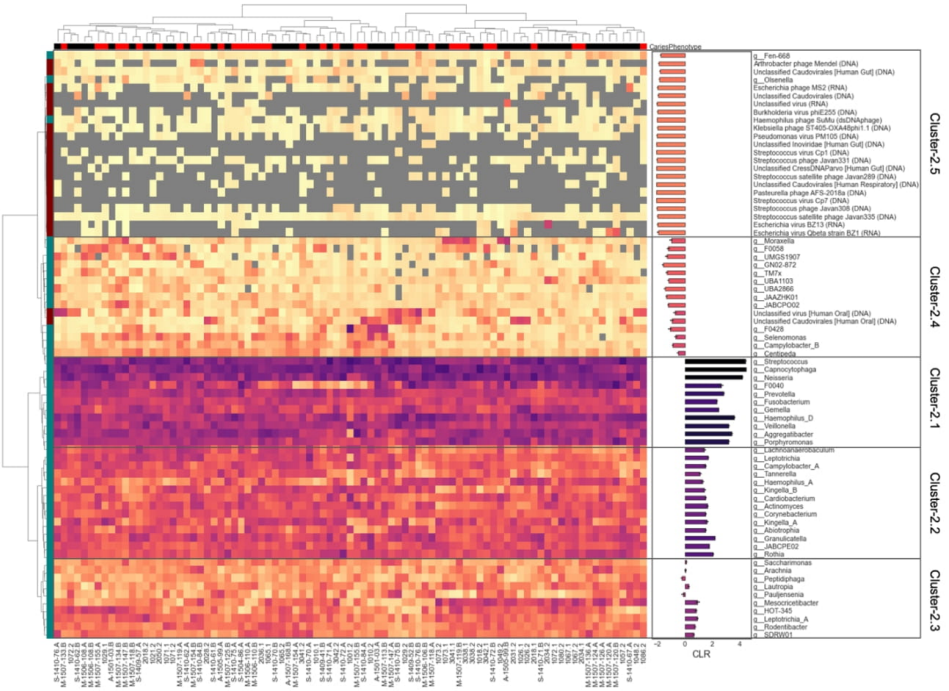


Fig. 2 – Taxonomic expression
(A) Center log-ratio transformed abundances of taxa from metatranscriptomics. Row colors represent bacterial (teal) or viral (maroon) MAGs while the column colors represent caries (red) or caries-free (black) phenotypes. Clustering was performed using Euclidean distance and Ward linkage.

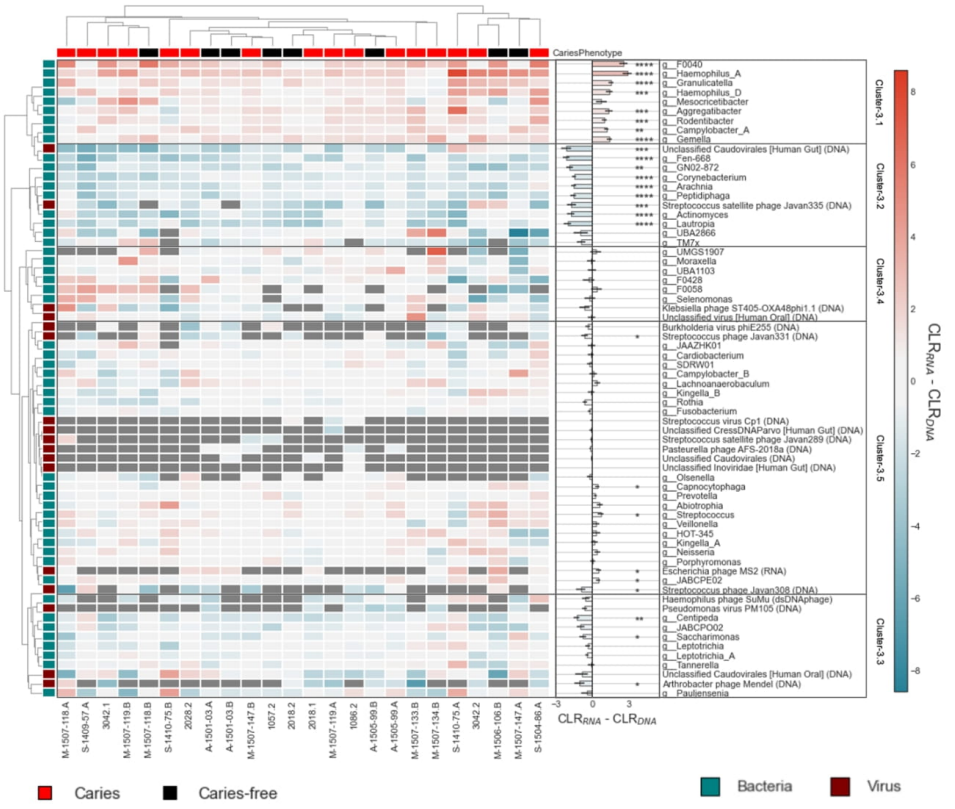


Fig. 3 – Taxonomic expression:abundance ratios.
Difference in CLR between metatranscriptomics expression (RNA) and metagenomics abundance (DNA) where values indicate higher expression relative to abundance and vice versa.

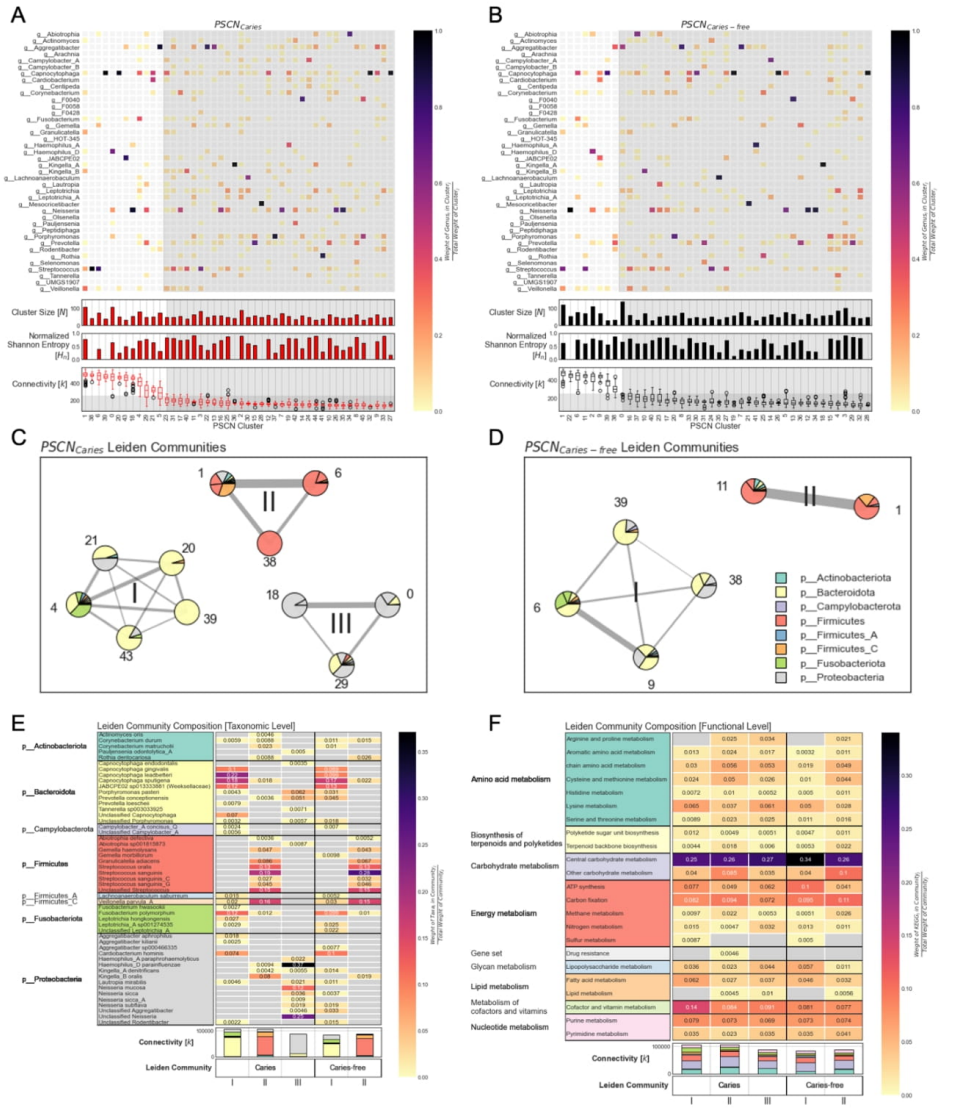


Fig. 4 – Connectivity-based community detection in PSCNs
Heatmap of clustered PSCNs for (A) caries and (B) caries-free phenotypes sorted by median cluster connectivity [*k*] in box-plot below with threshold for high-connectivity clusters set at 250 *k* in both PSCNs. Each *i,j* value in the heatmap represents the weight of genus *i* in cluster *j* divided by the total weight of cluster *j*; that is, the weighted proportion of each genus in each cluster. Leiden community detection algorithm applied to high-connectivity PSCN clusters for (C) caries and (D) caries-free phenotypes. Roman numerals indicate PSCN-specific Leiden communities for reference. Pie charts indicate proportion of genus weight in each Leiden community and colored by phyla. Clustering was performed using the distance

version of *rho* proportionality and Ward linkage. Heatmaps of Leiden Community connectivity (C,D) relative to taxonomy (E) and KEGG functional pathways (F) showing the connectivity of each grouping relative to the total connectivity in the community.

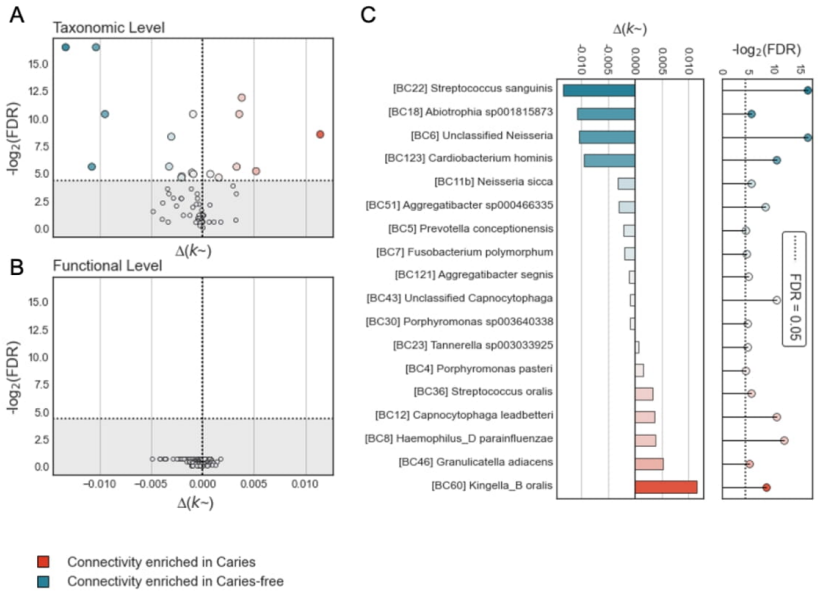


Fig. 5 – Comparing PSCNs with respect to taxonomic or functional levels
Volcano plots of Leiden community PGFCs from Fig. 4 showing change in scaled connectivity [$\Delta\tilde{k}$] and -$\log_2$(FDR) with respect to (A) taxonomic and (B) functional PGFC levels. (C) Sorted barchart of taxa with statistically different connectivities between caries and caries-free PSCNs. FDRs computed using Wilcoxon signed-rank test followed by Benjamini/Hochberg multiple hypothesis correction. Red represents an enrichment in connectivity in the caries PSCN with blue represents an enrichment in connectivity in the caries-free PSCN.

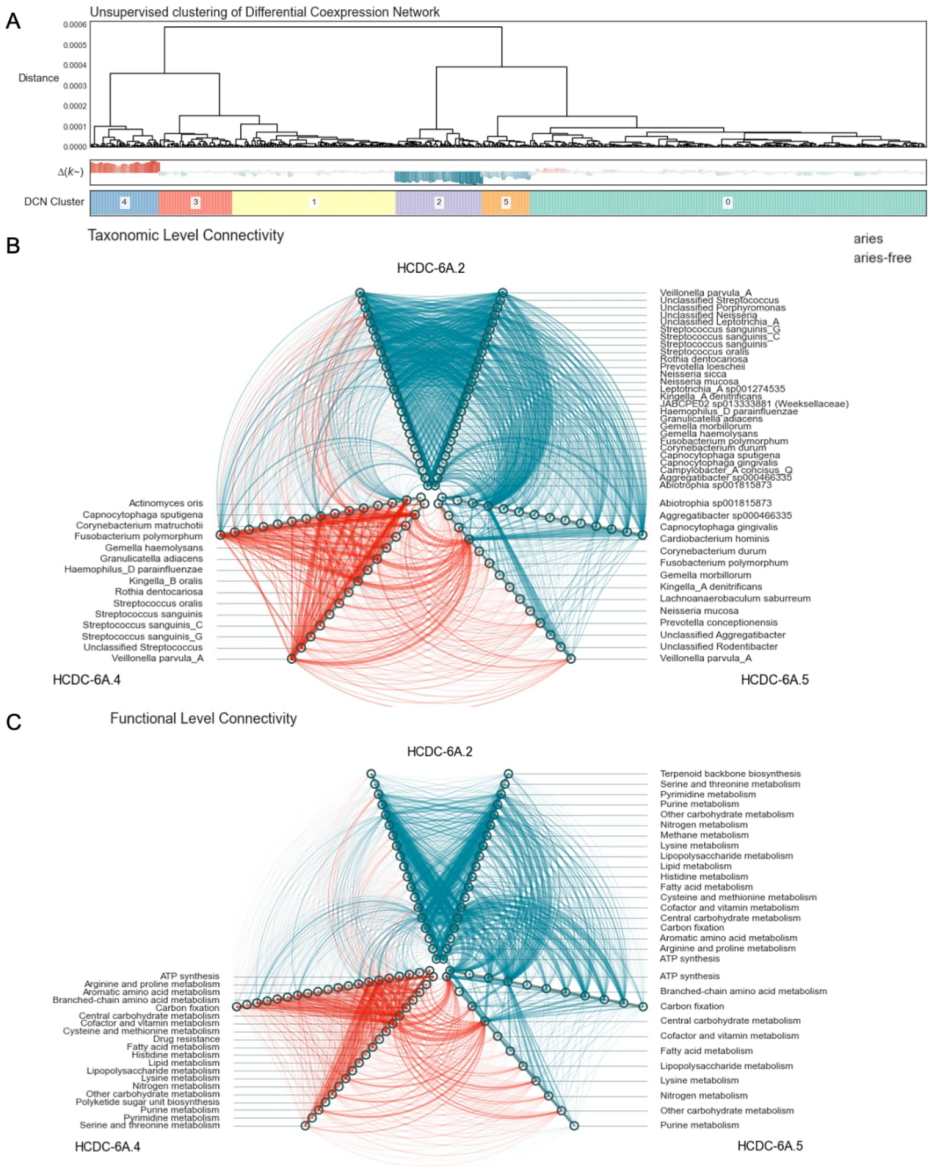


Fig. 6 – Differential network analysis between caries and caries-free PSCNs
(A) Hierarchical clustering of DCN using Leiden community PGFCs from Fig. 4. Barchart shows the differential connectivity [Δk^{-}] for PGFC nodes with positive (red) values indicating higher scaled connectivity

667 in caries PSCN and negative (blue) values indicating higher scaled connectivity in caries-free PSCN.
668 Colored panel on bottom shows DCN clusters sorted by the number of PGFCs in cluster with the largest
669 cluster being 0 and the smallest being 5. (B) Shows hive plot of taxonomic categories for DCN(Cluster-4),
670 DCN(Cluster-2), and DCN(Cluster-5) with red and blue edges following the scheme in (A). (C) Shows the
671 same hive plot in (B) but grouping PGFCs by higher-level KEGG categories instead of taxonomic
672 categories.

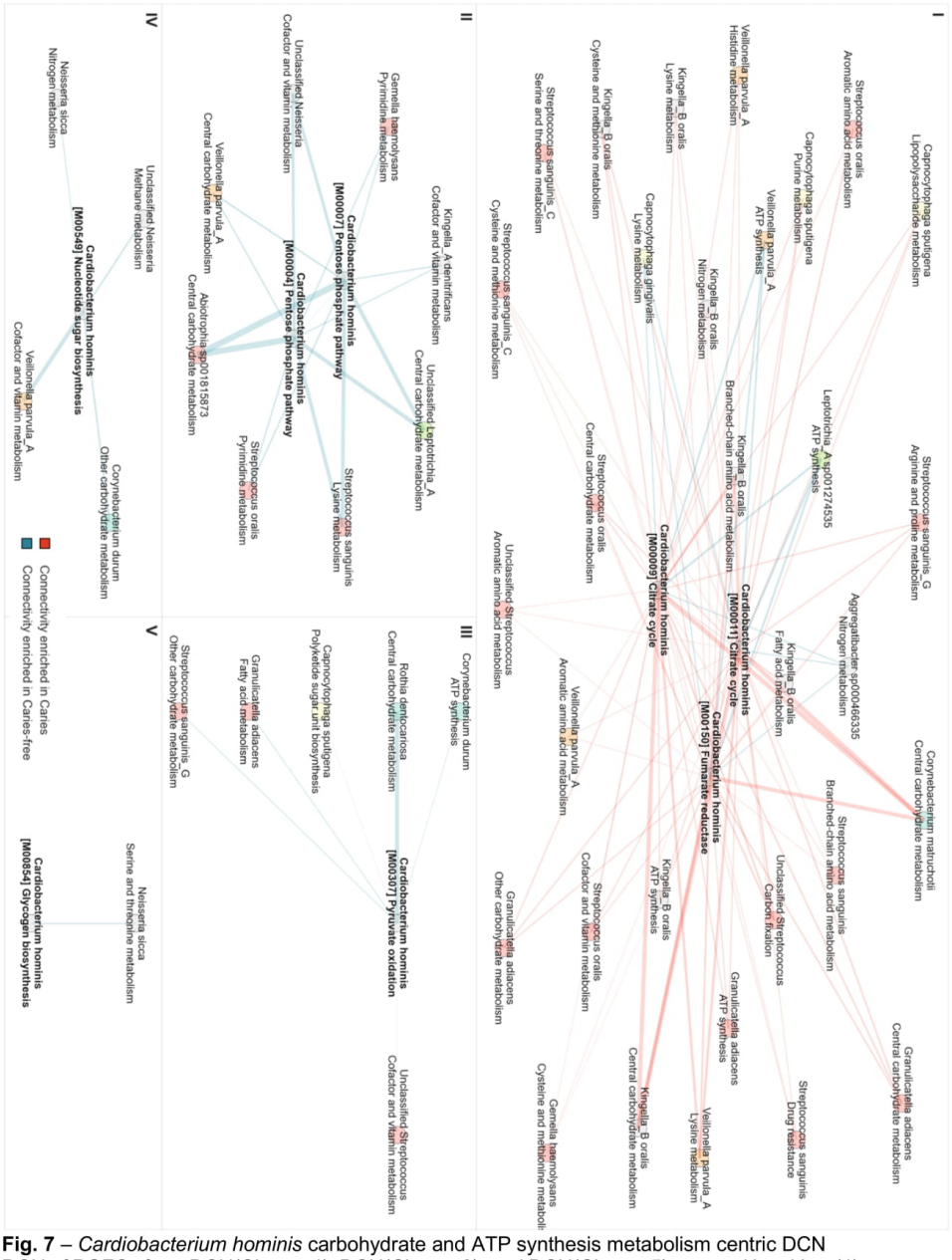
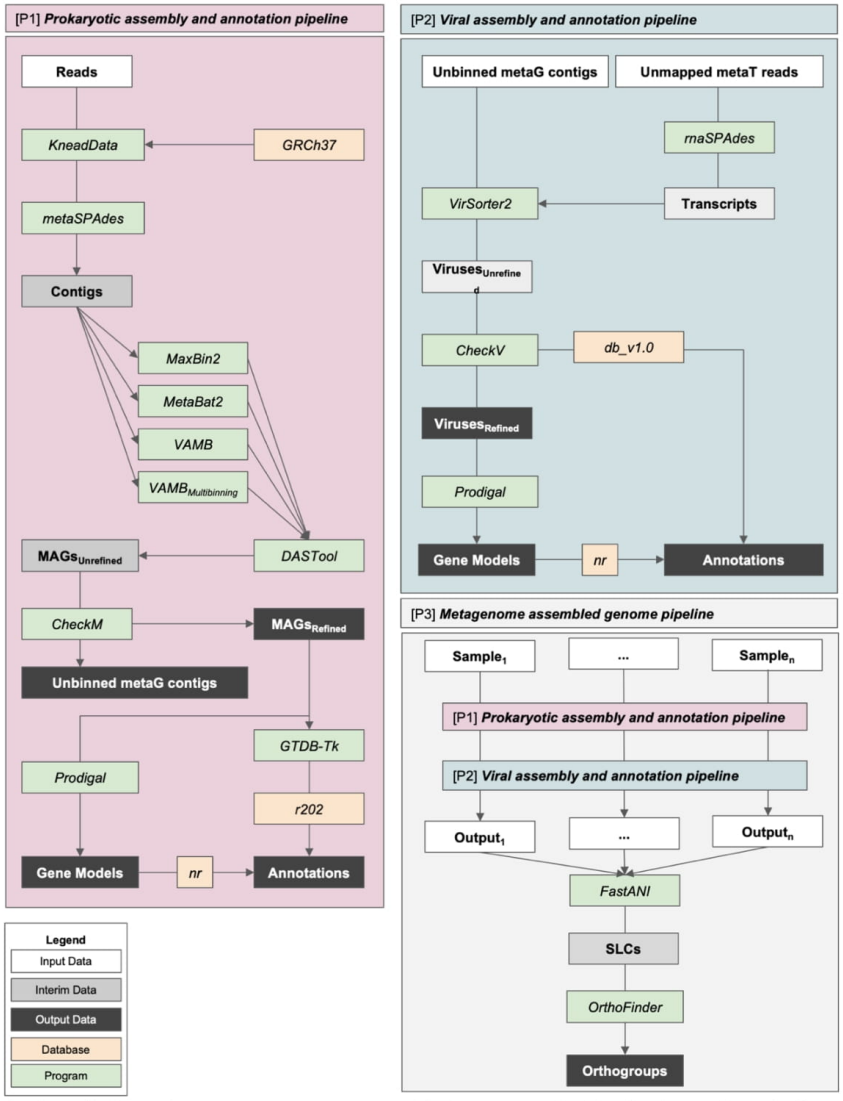


Fig. 7 – *Cardiobacterium hominis* carbohydrate and ATP synthesis metabolism centric DCN
DCN of PGFCs from DCN(Cluster-4), DCN(Cluster-2), and DCN(Cluster-5) grouped by either (1) taxonomy and higher-level KEGG categories for non-*C. hominis* PGFCs; and (2) *C. hominis* KEGG modules related

677 to carbohydrate and ATP synthesis metabolism. Edge weights indicate differential connectivity with positive
678 (red) edges indicating higher scaled connectivity in caries PSCN and negative (blue) edges indicating
679 higher scaled connectivity in caries-free PSCN.
680 Roman numerals indicate connected components in DCN (i.e. isolated subgraphs within the larger graph).
681
682 **Supplementary Material:**



683
684 **Fig. S1 – Schematic of metagenome assembled genome pipeline for bacteria and viruses**
685 (Left) Prokaryotic and (Right) viral assembly and annotation pipelines. (Bottom) Shows metagenomic
686 assembled genome pipeline.

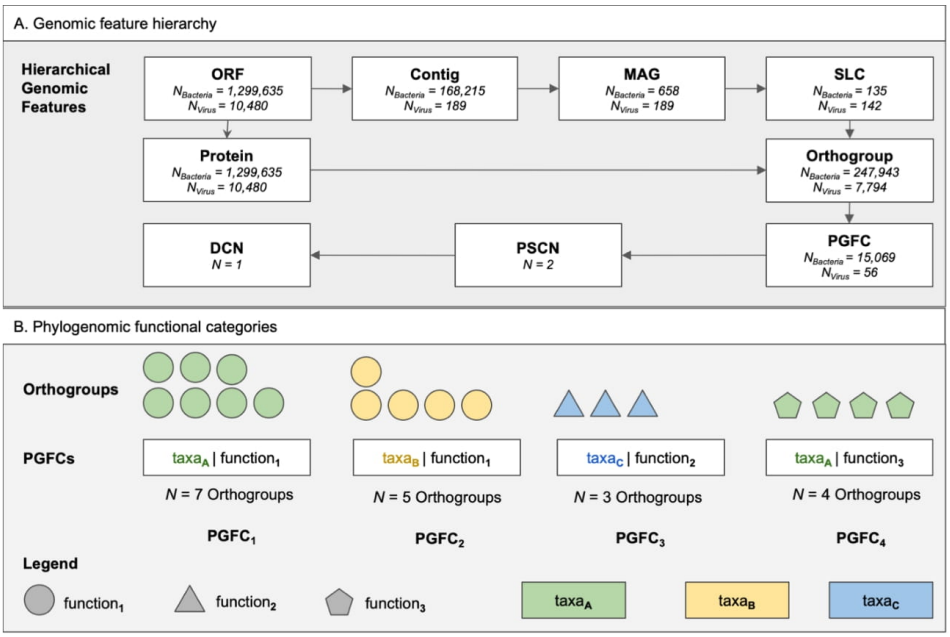


Fig. S2 – Genomic feature hierarchy and engineered features
FastANI clusters and orthogroups. (B) PGFC schematics showing taxonomic and functional categories created by grouping counts from orthogroups.

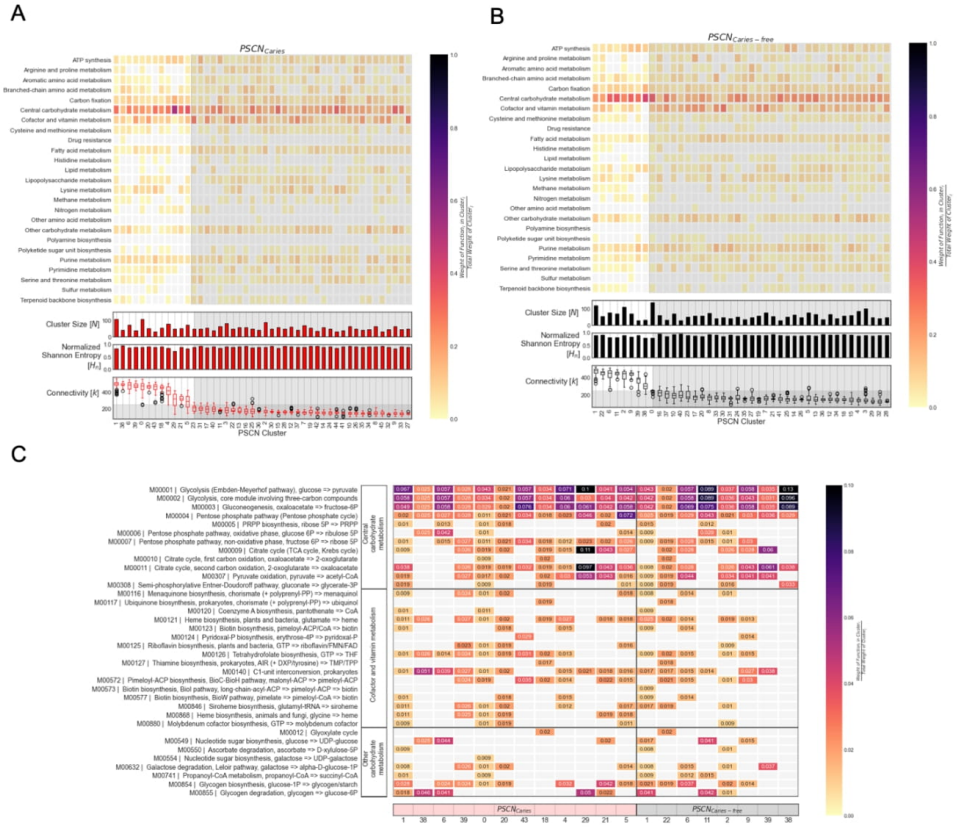


Fig. S3 – PSCN connectivity with respect to functional categories

Connectivity profiles for (A) caries and (B) caries-free PSCNs with respect to KEGG categories. (C) Proportion of connectivity for KEGG modules with respect to a PSCN clucster.

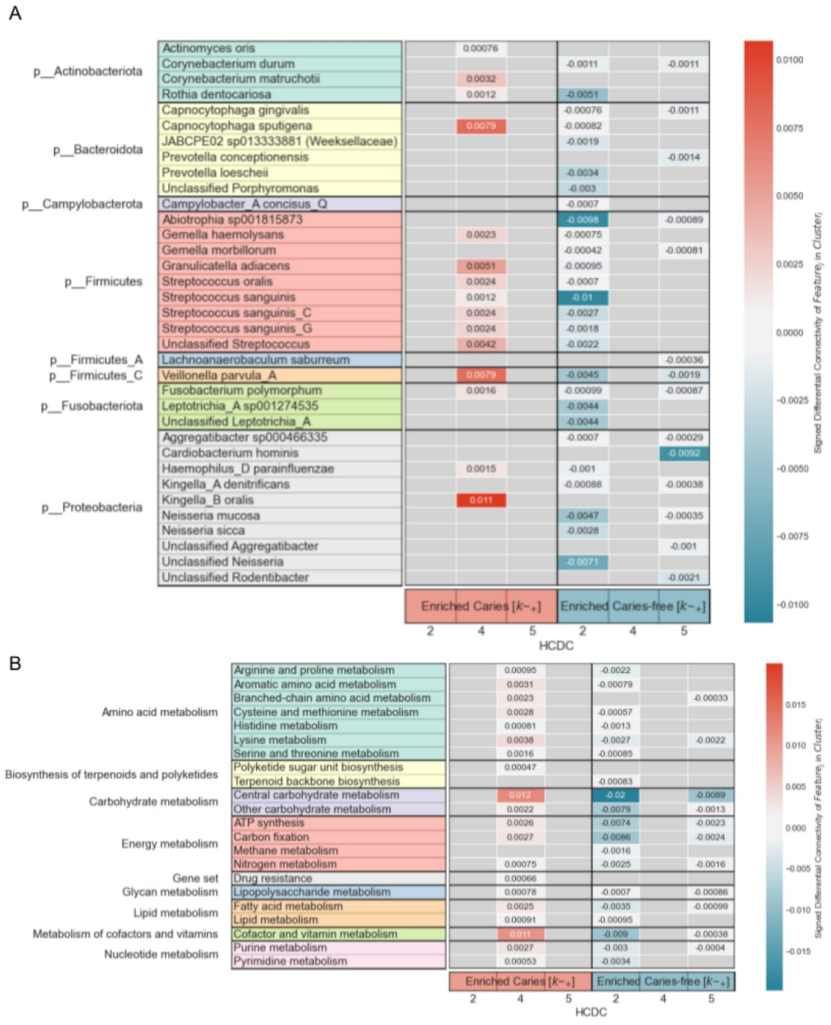


Fig. S4 – Taxonomic and functional membership with respect to HCDCs
Change in connectivity for each (A) phyla and (B) KEGG category with respect to HCDCs with (Left) connectivity enriched in caries and (Right) connectivity enriched in caries-free.

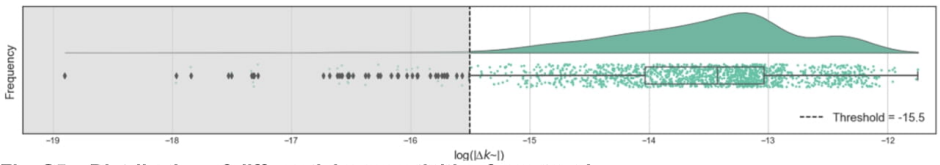


Fig. S5 – Distribution of differential connectivities for network
Distribution of differential connectivities for DCN in Fig. 7.

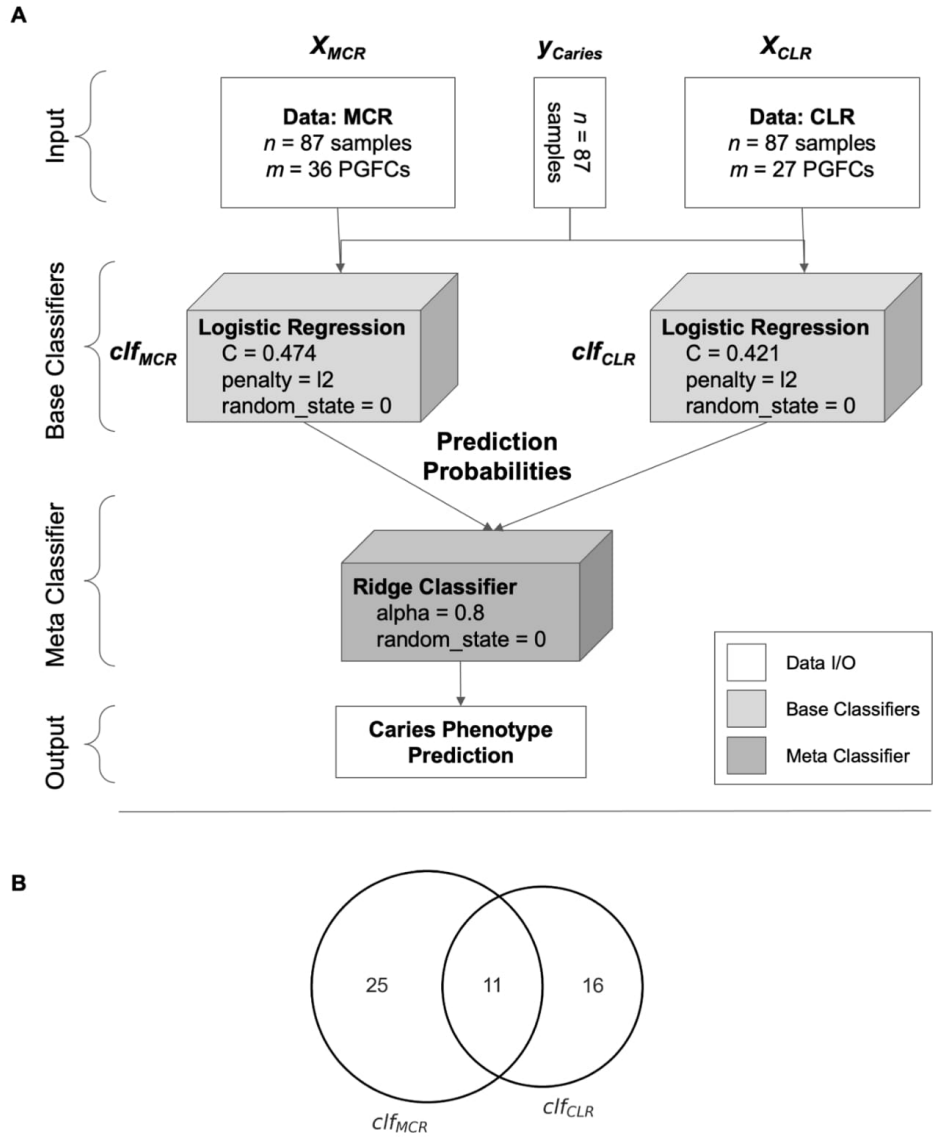


Fig. S6 – Mixed feature stacking classifier ensemble
(A) Schematic of stacking ensemble model with training data, base classifiers, meta classifier, and final prediction where *X, y,* and *clf* represent the feature matrix, target vector, and base classifier, respectively.

Feature set specific base models feed into a meta classifier which outputs the final prediction. MCR refers to module completion ratios of PGFCs and CLR refers to center log-ratio transformed abundances of PGFCs. Cubes indicate classification models with *sklearn* algorithm name in bold, followed by hyperparameter values indented (e.g., C, penalty, alpha, and random_state). (B) Venn diagram of PGFC features used to construct clf_{MCR} and clr_{CLR}.

Table S1 – Sample metadata and mapping to NCBI identifiers
Table S2 – MAG metadata including CheckM/V, classification, sample of origin, and identifiers
Table S3 – SLC metadata including taxonomy and genomes
Table S4 – SLC-specific orthogroup metadata
Table S5 – PGFC metadata and network prevalence
Table S6 – PSCN and DCN edge weights
Table S7 – Network clusters and communities using PGFCs as nodes
Table S8 – High-level network DCN edge weights
Table S9 – Features and weights for stacking classifier predictive model

Data Availability:
Metatranscriptomes and metagenomes were deposited in NCBI under BioProject PRJNA383868. Counts tables (ORF, orthogroup, MAG), genomes, gene models, and annotations are available on FigShare (doi: 10.6084/m9.figshare.18180614). Reproducible methods for feature engineering and network analysis are available at https://github.com/jolespin/ensemble_networkx.

Acronyms:
MAG – Metagenome Assembled Genome
SLC – Species-Level Cluster
CPR – Candidate Phyla Radiation
PSCN – Phenotype-Specific Coexpression Network
PGFC – Phylogenomic Functional Category
HCPC – High Connectivity PSCN Cluster
DCN – Differential Coexpression Network
HCDC – High Connectivity DCN Cluster
LFOCV – Leave Family Out Cross Validation

Author Contributions:
Conceptualization:
Josh L. Espinoza
Chris L. Dupont
Karen E. Nelson
Toby Hughes
Jeffrey M. Craig

Data Curation:
Josh L. Espinoza

Formal Analysis:
Josh L. Espinoza

Funding Acquisition:
Karen E. Nelson
Toby Hughes
Richard Saffery
Jeffrey M. Craig

Sample Collection:
Michelle Bockmann
Toby Hughes

Richard Saffery
Jeffrey M. Craig

DNA and RNA extractions:
Manolito G. Torralba
Claire Kuelbs

Methodology:
Josh L. Espinoza
Chris L. Dupont

Project Administration:
Chris L. Dupont
Karen E. Nelson

Resources:
Chris L. Dupont
Karen E. Nelson
Manolito G. Torralba

Software:
Josh L. Espinoza

Supervision:
Chris L. Dupont
Karen E. Nelson
Suren Singh

Validation:
Josh L. Espinoza
Chris L. Dupont

Visualization:
Josh L. Espinoza

Writing – Original Draft Preparation:
Josh L. Espinoza
Chris L. Dupont

Writing – Review & Editing:
Josh L. Espinoza
Chris L. Dupont
Pamela Leong

References:

Altman, N., & Krzywinski, M. (2018). The curse(s) of dimensionality. *Nature Methods*, *15*(6), 399–400. https://doi.org/10.1038/S41592-018-0019-X

Amat, S., Lantz, H., Munyaka, P. M., & Willing, B. P. (2020). Prevotella in Pigs: The Positive and Negative Associations with Production and Health. *Microorganisms*, *8*(10), 1–27. https://doi.org/10.3390/MICROORGANISMS8101584

BJ, K., E, Z., SM, H., JM, van der V., FH, S., RC, M., JM, ten C., & W, C. (2008). Pyrosequencing analysis of the oral microflora of healthy adults. *Journal of Dental Research*, *87*(11), 1016–1020. https://doi.org/10.1177/154405910808701104

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Chen, L., Collij, V., Jaeger, M., van den Munckhof, I. C. L., Vich Vila, A., Kurilshikov, A., Gacesa, R., Sinha, T., Oosting, M., Joosten, L. A. B., Rutten, J. H. W., Riksen, N. P., Xavier, R. J., Kuipers, F., Wijmenga, C., Zhernakova, A., Netea, M. G., Weersma, R. K., & Fu, J. (2020). Gut microbial co-abundance networks show specificity in inflammatory bowel disease and obesity. *Nature Communications 2020 11:1*, *11*(1), 1–12. https://doi.org/10.1038/s41467-020-17840-y

Cherkasov, S. V., Popova, L. Y., Vivtanenko, T. V., Demina, R. R., Khlopko, Y. A., Balkin, A. S., & Plotnikov, A. O. (2019). Oral microbiomes in children with asthma and dental caries. *Oral Diseases*, *25*(3), 898–910. https://doi.org/10.1111/ODI.13020

Dye, B. A., Li, X., & Beltran-Aguilar, E. D. (2012). Selected oral health indicators in the United States, 2005-2008. *NCHS Data Brief*, *96*, 1–8.

Erb, I., & Notredame, C. (2016). How should we measure proportionality on relative gene expression data? *Theory in Biosciences*, *135*(1–2), 21–36. https://doi.org/10.1007/s12064-015-0220-8

Espinoza, J. L., Dupont, C. L., O'Rourke, A., Beyhan, S., Morales, P., Spoering, A., Meyer, K. J., Chan, A. P., Choi, Y., Nierman, W. C., Lewis, K., & Nelson, K. E. (2021). Predicting antimicrobial mechanism-of-action from transcriptomes: A generalizable explainable artificial intelligence approach. *PLOS Computational Biology*, *17*(3), e1008857. https://doi.org/10.1371/journal.pcbi.1008857

Espinoza, J. L., Harkins, D. M., Torralba, M., Gomez, A., Highlander, S. K., Jones, M. B., Leong, P., Saffery, R., Bockmann, M., Kuelbs, C., Inman, J. M., Hughes, T., Craig, J. M., Nelson, K. E., & Dupont, C. L. (2018). Supragingival Plaque Microbiome Ecology and Functional Potential in the Context of Health and Disease. *MBio*, *9*(6). https://doi.org/10.1128/mBio.01631-18

Espinoza, J. L., Shah, N., Singh, S., Nelson, K. E., & Dupont, C. L. (2020). Applications of weighted association networks applied to compositional data in biology. *Environmental Microbiology*, *22*(8), 3020–3038. https://doi.org/10.1111/1462-2920.15091

Folke, C., Carpenter, S., Walker, B., Scheffer, M., Elmqvist, T., Gunderson, L., & Holling, C. S. (2004). Regime shifts, resilience, and biodiversity in ecosystem management. In *Annual Review of Ecology, Evolution, and Systematics* (Vol. 35, pp. 557–581). Annual Reviews. https://doi.org/10.1146/annurev.ecolsys.35.021103.105711

Freire, M., Moustafa, A., Harkins, D. M., Torralba, M. G., Zhang, Y., Leong, P., Saffery, R., Bockmann, M., Kuelbs, C., Hughes, T., Craig, J. M., & Nelson, K. E. (2020). Longitudinal Study of Oral Microbiome Variation in Twins. *Scientific Reports*, *10*(1), 7954. https://doi.org/10.1038/s41598-020-64747-1

Fuller, T. F., Ghazalpour, A., Aten, J. E., Drake, T. A., Lusis, A. J., & Horvath, S. (2007). Weighted gene coexpression network analysis strategies applied to mouse weight. *Mammalian Genome*, *18*(6–7), 463–472. https://doi.org/10.1007/s00335-007-9043-3

Gomez, A., Espinoza, J. L., Harkins, D. M., Leong, P., Saffery, R., Bockmann, M., Torralba, M., Kuelbs, C., Kodukula, R., Inman, J., Hughes, T., Craig, J. M., Highlander, S. K., Jones, M. B., Dupont, C. L., & Nelson, K. E. (2017). Host Genetic Control of the Oral Microbiome in Health and Disease. *Cell Host & Microbe*, *22*(3), 269-278.e3. https://doi.org/10.1016/j.chom.2017.08.013

Hsu, C.-L., Juan, H.-F., & Huang, H.-C. (2015). Functional Analysis and Characterization of Differential Coexpression Networks. *Scientific Reports*, *5*(1), 13295. https://doi.org/10.1038/srep13295

Humphrey, L. T., De Groote, I., Morales, J., Barton, N., Collcutt, S., Ramsey, C. B., & Bouzouggar, A. (2014). Earliest evidence for caries and exploitation of starchy plant foods in Pleistocene hunter-gatherers from Morocco. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(3), 954–959. https://doi.org/10.1073/pnas.1318176111

Jackson, M. A., Bonder, M. J., Kuncheva, Z., Zierer, J., Fu, J., Kurilshikov, A., Wijmenga, C., Zhernakova, A., Bell, J. T., Spector, T. D., & Steves, C. J. (2018). Detection of stable community structures within gut microbiota co-occurrence networks from different human populations. *PeerJ*, *6*(2). https://doi.org/10.7717/PEERJ.4303

James, S. L., Abate, D., Abate, K. H., Abay, S. M., Abbafati, C., Abbasi, N., Abbastabar, H., Abd-Allah, F., Abdela, J., Abdelalim, A., Abdollahpour, I., Abdulkader, R. S., Abebe, Z., Abera, S. F., Abil, O. Z., Abraha, H. N., Abu-Raddad, L. J., Abu-Rmeileh, N. M. E., Accrombessi, M. M. K., … Murray, C. J. L. (2018). Global, regional, and national incidence, prevalence, and years lived with disability for 354 Diseases and Injuries for 195 countries and territories, 1990-2017: A systematic analysis for the Global Burden of Disease Study 2017. *The Lancet*, *392*(10159), 1789–1858. https://doi.org/10.1016/S0140-6736(18)32279-7

Kanasi, E., Dewhirst, F. E., Chalmers, N. I., Kent, R., Jr., Moore, A., Hughes, C. V., Pradhan, N., Loo, C. Y., & Tanner, A. C. R. (2010). Clonal Analysis of the Microbiota of Severe Early Childhood Caries. *Caries Research*, *44*(5), 485. https://doi.org/10.1159/000320158

Kleinberg, I. (2002). A mixed-bacteria ecological approach to understanding the role of the oral bacteria in dental caries causation: an alternative to Streptococcus mutans and the specific-plaque hypothesis. *Critical Reviews in Oral Biology and Medicine : An Official Publication of the American Association of Oral Biologists*, *13*(2), 108–125.

Krzywinski, M., Birol, I., Jones, S. J., & Marra, M. A. (2012). Hive plots--rational approach to visualizing networks. *Briefings in Bioinformatics*, *13*(5), 627–644. https://doi.org/10.1093/bib/bbr069

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*(1), 559. https://doi.org/10.1186/1471-2105-9-559

Liljemark, W. F., Bloomquist, C. G., Uhl, L. A., Schaffer, E. M., Wolff, L. F., Pihlstrom, B. L., & Bandt, C. L. (1984). Distribution of oral Haemophilus species in dental plaque from a large adult population. *Infection and Immunity*, *46*(3), 778.

Lovell, D., Pawlowsky-Glahn, V., Egozcue, J. J., Marguerat, S., & Bähler, J. (2015). Proportionality: A Valid Alternative to Correlation for Relative Data. *PLOS*

906 *Computational Biology*, *11*(3), e1004075.
907 https://doi.org/10.1371/journal.pcbi.1004075
908 Marsh, P. D. (1994). Microbial Ecology of Dental Plaque and its Significance in Health
909 and Disease. *Advances in Dental Research*, *8*(2), 263–271.
910 https://doi.org/10.1177/08959374940080022001
911 Marsh, P. D., & Bradshaw, D. J. (1997). Physiological approaches to the control of oral
912 biofilms. In *Advances in dental research* (Vol. 11, Issue 1, pp. 176–185). Adv Dent
913 Res. https://doi.org/10.1177/08959374970110010901
914 McInnes, L., Healy, J., & Melville, J. (2018). *UMAP: Uniform Manifold Approximation*
915 *and Projection for Dimension Reduction*.
916 Morton, J. T., Marotz, C., Washburne, A., Silverman, J., Zaramela, L. S., Edlund, A.,
917 Zengler, K., & Knight, R. (2019). Establishing microbial composition measurement
918 standards with reference frames. *Nature Communications*, *10*(1), 1–11.
919 https://doi.org/10.1038/s41467-019-10656-5
920 Morton, J. T., Sanders, J., Quinn, R. A., McDonald, D., Gonzalez, A., Vázquez-Baeza,
921 Y., Navas-Molina, J. A., Song, S. J., Metcalf, J. L., Hyde, E. R., Lladser, M.,
922 Dorrestein, P. C., & Knight, R. (2017). Balance Trees Reveal Microbial Niche
923 Differentiation. *MSystems*, *2*(1), e00162-16.
924 https://doi.org/10.1128/mSystems.00162-16
925 N, U. T., V, A. C. M., E, U., I, W., US, J., H, J.-P., T, M., O, A., G, K., G, S., & E, N.
926 (2018). Performance of mass spectrometric identification of clinical Prevotella
927 species using the VITEK MS system: A prospective multi-center study. *Anaerobe*,
928 *54*, 205–209. https://doi.org/10.1016/J.ANAEROBE.2018.05.016
929 Nabwera, H. M., Espinoza, J. L., Worwui, A., Betts, M., Okoi, C., Sesay, A. K., Bancroft,
930 R., Agbla, S. C., Jarju, S., Bradbury, R. S., Colley, M., Jallow, A. T., Liu, J., Houpt,
931 E. R., Prentice, A. M., Antonio, M., Bernstein, R. M., Dupont, C. L., & Kwambana-
932 Adams, B. A. (2021). Interactions between fecal gut microbiome, enteric
933 pathogens, and energy regulating hormones among acutely malnourished rural
934 Gambian children. *EBioMedicine*, *73*, 103644.
935 https://doi.org/https://doi.org/10.1016/j.ebiom.2021.103644
936 Nyvad, B., & Takahashi, N. (2020). Integrated hypothesis of dental caries and
937 periodontal diseases. In *Journal of Oral Microbiology* (Vol. 12, Issue 1). Taylor and
938 Francis Ltd. https://doi.org/10.1080/20002297.2019.1710953
939 OECD. (2017). *Health at a Glance 2017: OECD Indicators* (Health at a Glance). OECD.
940 https://doi.org/10.1787/health_glance-2017-en
941 Paulson, J. N., Stine, O. C., Bravo, H. C., & Pop, M. (2013). Differential abundance
942 analysis for microbial marker-gene surveys. *Nature Methods*, *10*(12), 1200–1202.
943 https://doi.org/10.1038/nmeth.2658
944 Peres, M. A., Macpherson, L. M. D., Weyant, R. J., Daly, B., Venturelli, R., Mathur, M.
945 R., Listl, S., Celeste, R. K., Guarnizo-Herreño, C. C., Kearns, C., Benzian, H.,
946 Allison, P., & Watt, R. G. (2019). Oral diseases: a global public health challenge. In
947 *The Lancet* (Vol. 394, Issue 10194, pp. 249–260). Lancet Publishing Group.
948 https://doi.org/10.1016/S0140-6736(19)31146-8
949 Quinn, T. P., Erb, I., Richardson, M. F., & Crowley, T. M. (2018). Understanding
950 sequencing data as compositions: an outlook and review. *Bioinformatics (Oxford,*
951 *England)*, *34*(16), 2870–2878. https://doi.org/10.1093/bioinformatics/bty175

952 Quinn, T. P., Richardson, M. F., Lovell, D., & Crowley, T. M. (2017). Propr: An R-
953 package for Identifying Proportionally Abundant Features Using Compositional
954 Data Analysis. *Scientific Reports*, *7*(1), 1–9. https://doi.org/10.1038/s41598-017-
955 16520-0
956 Shaiber, A., & Eren, A. M. (2019). Composite metagenome-assembled genomes reduce
957 the quality of public genome repositories. *MBio*, *10*(3).
958 https://doi.org/10.1128/MBIO.00725-19
959 Silverman, J. D., Roche, K., Mukherjee, S., & David, L. A. (2020). Naught all zeros in
960 sequence count data are the same. *Computational and Structural Biotechnology*
961 *Journal*, *18*, 2789–2798. https://doi.org/10.1016/J.CSBJ.2020.09.014
962 Smith, S. R., Gillard, J. T. F., Kustka, A. B., McCrow, J. P., Badger, J. H., Zheng, H.,
963 New, A. M., Dupont, C. L., Obata, T., Fernie, A. R., & Allen, A. E. (2016).
964 Transcriptional Orchestration of the Global Cellular Response of a Model Pennate
965 Diatom to Diel Light Cycling under Iron Limitation. *PLOS Genetics*, *12*(12),
966 e1006490. https://doi.org/10.1371/JOURNAL.PGEN.1006490
967 Takahashi, N., & Nyvad, B. (2011). The role of bacteria in the caries process: ecological
968 perspectives. *J Dent Res*, *90*(3), 294–303.
969 https://doi.org/10.1177/0022034510379602
970 Takahashi, N., Washio, J., & Mayanagi, G. (2010). Metabolomics of supragingival
971 plaque and oral bacteria. *Journal of Dental Research*, *89*(12), 1383–1388.
972 https://doi.org/10.1177/0022034510377792
973 Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden:
974 guaranteeing well-connected communities. *Scientific Reports*, *9*(1), 1–12.
975 https://doi.org/10.1038/s41598-019-41695-z
976 Van Der Maaten, L. (2014). Accelerating t-SNE using Tree-Based Algorithms. *Journal of*
977 *Machine Learning Research*, *15*, 1–21.
978 Wilson, S. J., Wilkins, A. D., Lin, C. H., Lua, R. C., & Lichtarge, O. (2017). Discovery of
979 functional and disease pathways by community detection in protein-protein
980 interaction networks. *Pacific Symposium on Biocomputing*, *0*(212679), 336–347.
981 https://doi.org/10.1142/9789813207813_0032
982 XH, Y., & M, R. (2021). Multi-label classification and label dependence in in silico
983 toxicity prediction. *Toxicology in Vitro : An International Journal Published in*
984 *Association with BIBRA*, *74*. https://doi.org/10.1016/J.TIV.2021.105157
985 Yang, X., He, L., Yan, S., Chen, X., & Que, G. (2021). The impact of caries status on
986 supragingival plaque and salivary microbiome in children with mixed dentition: a
987 cross-sectional survey. *BMC Oral Health 2021 21:1*, *21*(1), 1–13.
988 https://doi.org/10.1186/S12903-021-01683-0
989 Zaura, E., Keijser, B. J., Huse, S. M., & Crielaard, W. (2009). Defining the healthy “core
990 microbiome” of oral microbial communities. *BMC Microbiology*, *9*, 259.
991 https://doi.org/10.1186/1471-2180-9-259
992 Zheng, F., Zhang, S., Churas, C., Pratt, D., Bahar, I., & Ideker, T. (2021). HiDeF:
993 identifying persistent structures in multiscale ‘omics data. *Genome Biology 2021*
994 *22:1*, *22*(1), 1–15. https://doi.org/10.1186/S13059-020-02228-4
995

Supplementary Methods:

Assembly, classification, dereplication for metagenome assembled genomes (bacteria and viruses)

We pursued a metapantranscriptomics approach by assembling and binning each metagenome separately, performing species-level clustering, and a within SLC orthology analysis (Fig. S2A). The primary aim of this research is to investigate inferred interactions within the supragingival plaque microbiome and how these interactions are associated with either caries or caries-free phenotypes. With this scope in mind, we analyzed only the conserved regions of each SLC and, when mapping reads, produced a denser matrix that is better suited for network analysis where we aim to identify core patterns. While focusing on these core ecological features, we simultaneously provide high quality and biologically relevant MAGs for other researchers to perform pangenomics analysis.

Our metagenomics dataset contains 88 subjects and were characterized in our previous work using a coassembly approach (Espinoza et al., 2018). Using updated state-of-the-art consensus methodologies, we have revisited this dataset to provide higher quality MAGs and annotations resulting in more accurate biological interpretations. Each of these 88 processed metagenomic samples were assembled on a per sample basis using *SPAdes* v3.15.2 (*metaSPAdes* mode (Nurk et al., 2017) and binned using several binning algorithms including: 1) *MaxBin2* v2.2.7 (Wu et al., 2016), 2) *MetaBAT2* v2.15 (Kang et al., 2019), 3) *VAMB* v3.0.2 (Nissen et al., 2021), and 4) *VAMB* using "multi-split binning" which uses all samples together. The binning assignments were refined using *DAS Tool* v1.1.2 (Sieber et al., 2018) yielding 833 MAGs that were classified using *GTDB-Tk* v1.5.0 (reference database: r202 (Chaumeil et al., 2020)) indicating all MAGs were bacteria and none were archaea. The 883 bacteria MAGs were assessed using the *lineage_wf* pipeline of *CheckM* v1.1.3 (Parks et al., 2015) and 658 of which were of at least medium quality according to the guidelines established by the Genomic Standard Consortium (>50% completeness and <10% contamination; (Bowers et al., 2017) which are referred to here after as $Assembly_{Bacteria}$. For the *Patescibacteria* CPR, we used the 43 marker genes (cpr_43_markers.hmm) propose by Brown *et al.* 2015 (Brown et al., 2015). These 124 of the 658 MAGs had no confident species-level classification from *GTDB-Tk* and indicate novel species from known genera.

Contigs, both binned and unbinned, that were not included in $Assembly_{Bacteria}$ were input into *VirSorter2* v2.2.2 (--include-groups dsDNAphage,NCLDV,ssDNA,lavidaviridae; (Guo et al., 2021)) and quality assessed using *CheckV* v0.8.1 (reference database: v1.0; (Nayfach et al., 2020)) yielding 179 DNA viruses of at least medium quality; referred to as $Assembly_{VirusDNA}$. The recommended cutoffs in *CheckV* for considering viral contigs are the following: 1) number of viral genes > 0; 2) the number of viral genes ≥ 5 times the number of host genes; 3) completeness ≥ 50%; 4) *checkv_quality* and *miuvig_quality* are above medium quality (Nayfach, 2021). Reads from the 91 metatranscriptomic samples that did not map to any of the metagenomic contigs (both binned and unbinned) were assembled using *SPAdes* (*rnaSPAdes* mode) on a per sample basis. These transcript assemblies were input into *VirSorter2* (--include-groups RNA) and assessed for quality using *CheckV* yielding 10 RNA viruses; referred to as $Assembly_{VirusRNA}$. Classifications for viruses were assigned using the *aai_top_hit* field from *completeness.tsv* where RefSeq identifiers were propagated and DTR_[XXXXXX] identifiers were extrapolated from the genome_db/checkv_circular.tsv file of the *CheckV* database. Only non-proviruses were used in downstream analysis to prevent host bias.

Dereplication of redundant species were clustered using *FastANI* v1.32 (Jain et al., 2018) with a cutoff of ≥ 95% ANI, as recommended by authors, and connected components were determined using *NetworkX* v2.5 (Hagberg et al., 2008). FastANI clustering was performed separately on the following: 1) $Assembly_{Bacteria}$ resulting in 135 unique clusters, prefixed by BC, with 64 of the clusters representing singleton clusters of size 1; and 2) $Assembly_{VirusDNA}$ + $Assembly_{VirusRNA}$ resulting in 142 unique clusters, prefixed by VC, with 113 singleton clusters. Manual curation of *FastANI* defined clusters was performed when the *GTDB-Tk* classifications were clearly split (i.e. *FastANI* clusters: BC0, BC11, and BC25). These *FastANI* clusters are referred to through this study as species-level clusters (SLC).

The 289,342 contigs from $Assembly_{Bacteria}$ (658 bacterial MAGs), $Assembly_{VirusDNA}$ (179 DNA viruses), and $Assembly_{VirusRNA}$ (10 RNA viruses) were combined into the $Assembly_{Catalogue}$ that is used for all read mapping, gene modeling, and statistical analysis. A flowchart for bacterial, viral, and orthogroup pipelines are available in Fig. S1.

Gene models, functional annotation, and orthology analysis

Gene models were created for *Assembly*$_{Catalogue}$ using *Prodigal* v2.6.3 (--meta mode; (Hyatt et al., 2010)) generating putative proteins and GFF files used for annotations and read mapping. The putative proteins were annotated using both *Diamond* v2.0.8 (Buchfink et al., 2014) against RefSeq's non-redundant database (accessed 2020.04.02) KOFAMSCAN v1.3.0 (Aramaki et al., 2020).

Orthology analysis was performed for each putative proteins in each *FastANI* cluster using OrthoFinder v2.5.2 (Emms & Kelly, 2019). For singleton *FastANI* clusters, dummy genomes were provided to yield a paralog-style analysis to stay consistent with the orthogroup analysis where intra-genome orthogroups are permitted. Consensus annotations for were determined for each orthogroup using the annotations from Diamond and KOFAMSCAN.

Read preprocessing and mapping to genome catalogue

HISAT2 v2.2.1 (Kim et al., 2019) was used for all read mapping to *Assembly*$_{Catalogue}$ for both metagenomics and metatranscriptomics datasets. Read counts tables were constructed with *featureCounts* v2.0.1 (Liao et al., 2014) using the bam files generated from *HISAT2,* the *Assembly*$_{Catalogue}$ fasta, and the gene models defined using *Prodigal*. Counts for FastANI clusters or MAGs were computed by summing contig-level read counts using a SAF file generated from *Assembly*$_{Catalogue}$.

Reads were preprocessed using *KneadData v0.6.1* by quality trimming using *Trimmomatic* v0.39 (Bolger et al., 2014), aligning reads to the human genome (GRCh37.p4) via *Bowtie2* v2.4.3 (Langmead & Salzberg, 2012), and removing aligned reads for assembly and mapping. The metagenomics and metatranscriptomics datasets contained reads with a post-processed sequencing depth of 2,946,448 ± 1,432,642 reads and 5,963,060 ± 1,726,024 reads, respectively. The alignment rate to *Assembly*$_{Catalogue}$ for metagenomics was 73.0 ± 7% and 69.7 ± 5.9% for metagenomics and metatranscriptomics datasets, respectively. The missing alignments are due to the fact that we only mapped to the highest quality bacterial MAGs and viral contigs.

Phylogenomic functional categories

Within our dataset, there are 255,737 SLC-specific orthogroups which would result in ~32.7 billion non-redundant connections in a fully-connected coexpression network; an insurmountable dataset for exploratory analysis on most modern machines. Instead of using draconian filtering thresholds of orthogroups, we sought to devise a feature engineering technique that would allow seamless transitions from read ↔ ORF/orthogroup ↔ contig/MAG/SLC ↔ engineered feature using custom taxonomy fields and functional assignments (e.g. KEGG, MetaCyc, PFAM).

With this in mind, we present the *PhyloGenomic Functional Category* (PGFC) as a supervised microbiome feature engineering method that can be used for high-level statistical analysis and unpacked back into individual orthogroups (or ORFs) unlike dimensionality reduction methods such as PC[o]A, *t*-SNE, or UMAP. PGFCs essentially group low-level features, orthogroups in this context, by a taxonomic unit (SLC) and a functional unit (KEGG module) (Fig. 3B, Table S5); similar, but not identical, to *HUMAnN* (Beghini et al., 2021; Franzosa et al., 2018) which does not allow the flexibility for custom low-level features from *de-novo* meta-omics. Another similar approach is the amalgram where compositions can either have exclusive or non-exclusive mappings between the original feature and engineered feature (Quinn & Erb, 2020). However, these engineered features cannot be collapsed and expanded with respect to predefined categories such as taxonomy and metabolism so will not be explored in this study.

PGFCs are composite features that group metabolic functional information with genome-resolved taxonomy assignments and were created by grouping all the orthogroups that had KEGG orthology, defined via KOFAMSCAN, and extending the grouping up the hierarchy to modules with respect to taxonomy. Taxonomy for PGFCs were assigned to the *FastANI* cluster of origin. PGFCs were implemented using our *EnsembleNetworkX* v2021.06.24 Python package (Espinoza, 2020b; Nabwera et al., 2021) via the *CategoricalEngineeredFeature* class.

We implemented strict filtering scheme to identify only the most robust patterns. MCRs were calculated from the KEGG orthologs defined by KOFAMSCAN using *MicrobeAnnotater* v2.0.5 (Ruiz-Perez et al.,

2021). PGFCs were removed if they did not have MCR > 50% in at least 20 samples resulting in 2,478 PGFCS representing 89 bacterial taxonomic units and 113 functional units from 8554 orthogroups. The rationale behind using such a stringent threshold is that many downstream associations may be misleading if only a small number of enzymes are represented in the module.

Our unfiltered PGFC dataset contains 15,125 PGFCs incorporating 171 SLCs and 267 KEGG modules. Although this dataset was relatively dense, many of the KEGG modules were incomplete with respect to a SLC in a sample. For increased biological interpretability, we filtered out PGFCs with KEGG modules that were largely incomplete, MCR < 0.5, to identify patterns with higher confidence amongst these high-level features; a feature, to our knowledge, is not implemented in similar methodologies. Our MCR-filtered PGFC dataset contained 2,478 PGFCs representing 89 taxonomic units and 113 functional units from 8,554 orthogroups; all of which are from bacterial SLCs. In orthogroup-space, these features would amount to ~37 million non-redundant connections but ~3 million% in PGFC-space effectively compressing the information content by ~92%; making prototyping and data exploration tractable on modern compute machines.

The PGFC counts were normalized via prevalence by scaling the counts by the total number of ORFs in each sample and in each PGFC. Our dataset prior to preprocessing produced 15,125 PGFCs representing 171 bacterial and viral taxonomic units and 267 functional units from 24,640 orthogroups.

Differential expression analysis

We tested several differential abundance packages including ALDEx2 v1.12.0 (Fernandes et al., 2014; Quinn et al., 2018) and ANCOM v0.5.6 (Biocore, 2020; Mandal et al., 2015). These analyses were performed at 2 levels: 1) summing contig counts per SLC as a measure of total taxonomic expression in a sample; and 2) the preprocessed PGFC feature table used as input in network analysis. For ALDEx2, we tried Welch's *t*-test, Wilcoxon test, Kruskal-Wallace, and GLM tests with "iqlr", "zero", and "all" denominators. For ANCOM, we tried alpha 0.5, tau 0.02, and theta 0.1 parameters.

Typically, expression:abundance ratios are calculated using log fold-change of relative abundances. For consistency with our compositional approach, differential RNA and DNA abundance were calculated by taking the difference in CLR between biosamples with statistical significance calculated via Mannwhitney U-test and Benjamini/Hochberg adjusted p-values. As the CLR transform is in log-space, the difference between RNA and DNA values are akin to log fold-change.

Ensemble networks and connectivity measurements

Ensemble coexpression networks were designed to robustly model associations within a system and differences in associations between systems such as phenotypes. In coexpression networks, it is assumed that weighted associations between features estimate biologically meaningful interactions within a system. Instead of creating a single cross-sectional network using all of the samples at once, we implement ensemble networks by bootstrapping samples, with or without repetition, and calculating a distribution of weighted associations as the basis for each edge in the networks. Ensemble machine-learning approaches have higher performance than singular methods (Opitz & Maclin, 1999) and boosting has been shown to resistant to outliers (Salibian-Barrera & Zamar, 2002)by minimizing the influence of individual samples and, in the context of networks, performed by representing edges as a distribution rather than a singular value.

For pairwise connectivity measurements, we implemented *rho* proportionality (Erb & Notredame, 2016; Espinoza, 2020a; Lovell et al., 2015; Quinn et al., 2017) which is a compositionally-aware association matrix akin to correlation. *Rho* proportionality is in the range [-1,1] where -1 means inversely proportional and 1 is perfectly proportional. Edge connectivity refers to the median *rho* proportionality associations from our ensemble networks and serve as the foundation for all higher-level connectivity measurements such as node connectivity, intra/inter-community, etc. For interpretability, we analyze only the positive associations as these have the most direct biological meaning.

Node connectivity is implemented as weighted-degree and is computed by summing all weighted edges connected to a particular node. Similar to the node connectivity calculations, we also investigate connectivity with respect to other higher-level groupings such as taxonomy, cluster, or metabolic function

by summing all of the non-self connectivities within that grouping. Differential connectivity is in reference to the caries-free cohort so that a positive DCN connectivity shows an enrichment in connectivity in the caries cohort while a negative connectivity shows a depletion in connectivity within the caries cohort.

Connectivity in this study is represented as either unscaled connectivity [k] or scaled connectivity [k^{-}] which is computed by scaling to total network connectivity in that the sum of scaled connectivities equals 1 in a network. Scaled connectivity is performed at the level of edges, that is all summed edge connections equal 1, and are useful when comparing networks with different total connectivity.

Phenotype-specific coexpression networks

In this study, we investigate differences in connectivity between caries and caries-free PSCNs using PGFCs as nodes and *rho* proportionality as the metric for associations (N_{nodes}=2,478; N_{edges}=3,069,003). These associations are interpreted as inferred interactions with insight into potential phenotype-specific metabolism coupled to taxonomy. Our ensemble coexpression network analysis is derived from the following network states: (1) caries cohort coexpression; and (2) caries-free cohort coexpression. After preprocessing as described previously, we constructed PSCNs using the following operations: (1) normalize the PGFC counts by dividing the summed counts by number of detected orthogroups in each sample relative to the PGFC grouping; (2) subset the data by phenotype; (3a) pseudo-random selection of 90% of the samples, (3b) pairwise *rho* proportionality for compositionally-aware associations among PGFCs within subject subset; (3c) stack non-redundant edge weights in array; (3d) repeat steps 3a-3c for 500 iterations with consecutive random seeds for reproducibility; (4) compute the median and median absolute deviation for each edge distribution as a representative edge weight; (5) split the positive associations ($\boldsymbol{A}_{+}$) and negative associations ($\boldsymbol{A}_{-}$) separately where $\boldsymbol{A}$ represents the association matrix; and (6) use the adjacency matrices for weighted edges to build fully-connected networks via *NetworkX v2.4*. For interpretability, only the positive associations were used in downstream network analysis. Normalized entropy for measuring network cluster homogeneity was computed using the following equation: $H_n(p) = -\sum_i \frac{p_i \log_2 p_i}{\log_2 n}$ where n is the number genera in the cluster and p_i is the weighted proportion of genus i in the cluster.

Differential network analysis

Our approach for quantifying the differences between network states is implemented using the following network types: (1) phenotype-specific coexpression networks (PSCN); and (2) differential coexpression networks (DCN). For PSCNs, we have networks that share the same nodes but differ in their edge weights. For instance, PGFC-level PSCNs have 2,478 PGFC nodes for both caries and caries-free systems. This property allows us to easily create DCNs from PSCNs by taking the difference between adjacency networks as follows: $PSCN_{Caries} - PSCN_{Caries\text{-}free} = DCN$.

We measure whether groupings such as taxonomy, metabolic function, and cluster have statistically different connectivities between caries and caries-free PSCNs via Wilcoxon signed rank-tests (*SciPy v1.4.1,* (Virtanen et al., 2020)). For Wilcoxon signed-rank tests, only groups that had n > 20 PGFCs were tested as recommended by the *SciPy* documentation. Benjamini/Hochberg multiple test correction was used to compute FDR values from Wilcoxon signed-rank test p-values.

Unsupervised network clustering and community detection

Many unsupervised methods can use a precomputed distance matrix as input. To seamlessly feed our networks into these unsupervised methods, we converted the network adjacency matrices to dissimilarity networks and used these as our precomputed distance matrix. As mentioned, we split our networks into positive and negative networks for downstream analysis. However, we used signed associations for clustering to maximize the information content used for the most informative clustering. To transform our signed association into a dissimilarity matrix that preserves all information content in the original association matrix, we used the formula $1 - rho_{+}$ which is an adaptation of correlation distance but applied to *rho* proportionality.

For unsupervised clustering, we use the following methods: (1) agglomerative hierarchical clustering of the dissimilarity network with ward linkage (*SciPy v1.4.1,* (Virtanen et al., 2020)) with hybrid methods for cutting

dendrograms (*dynamicTreeCut v1.63.1*); and (2) ensemble Leiden community detection of networks (Traag et al., 2019). Community detection does not necessarily reference *bona fide* microbial communities but more specifically communities of nodes within a network. Ensemble Leiden community detection was performed by running the Leiden algorithm (*leidenalg v0.8.3*) using 1000 different random seeds and grouping nodes that were in the same community for all iterations.

Agglomerative hierarchical clustering-based clusters are named by their respective figure and integer label. For example, Cluster-2.5 refers to cluster 5 of Fig. 2, HCPC-4A.1 refers to high connectivity PSCN cluster 1 of Fig. 4 panel A, and HCDC-6A.2 refers to high connectivity differential connectivity cluster 2 of Fig. 6 panel A. Leiden communities are named by their respective figure, panel, and roman numeral. For example, Community-4C.I is Leiden community I of Fig. 4 panel C.

Feature selection and predictive modeling

The *Clairvoyance* feature selection algorithm (Espinoza et al., 2021) was used to identify features that could discriminate caries phenotypes. The feature set used as input into the feature selection algorithm was the 212 PGFCs used in the DCN analysis. Of this feature set, two separate feature representations were used: 1) MCR; and 2) CLR where MCR refers to module completion ratios of PGFCs and CLR refers to center log-ratio transformed abundances. This implementation of stacking uses custom feature sets for each base classifier whose prediction probabilities are modeled using a meta classifier (Fig. S6, Table S9). These feature sets and their respective base classifiers were combined into a stacking classifier implemented in *sklearn v0.22.2* composed of clf_{MCR} (LogisticRegression(C=0.474, penalty=l2) and clf_{CLR} (LogisticRegression(C=0.421, penalty=l2) where *clf* refers to a base estimator whose output probabilities are used as input into a meta-classifier (RidgeClassifier(alpha=0.8)); C, penalty, and alpha refer to algorithm hyperparameters. The evaluation strategy used was Leave Family Out Cross Validation (LFOCV) accuracy, where family refers to the unique family identifier that corresponds to each twin or triplet pair, simulating predictive performance on new patients.

Visualization

Traditional network plots were implemented using a combination of *NetworkX v2.4* for plotting with *graphviz_layout* for graphical layout and *Matplotlib* as the backend. We use the *neato* layout algorithm from *graphviz* which creates virtual physical models and runs an iterative solver to find low energy configurations. Hive plots were implemented using the *Hive* class from *HiveNetworkX v2021.05.18* (Espinoza, 2020c; Krzywinski et al., 2012). Dendrograms were implemented using the *Agglomerative* class from *soothsayer v2021.05.18* (Espinoza, 2019; Espinoza et al., 2021). Heatmaps and clustermaps were implemented using *seaborn v0.9*. All other figures were generated with *Matplotlib v3.2.3* in *Python v3.8.2* unless specifically noted otherwise

References:

Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., & Ogata, H. (2020). KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics*, *36*(7), 2251–2252. https://academic.oup.com/bioinformatics/article/36/7/2251/5631907

Beghini, F., McIver, L. J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., Mailyan, A., Manghi, P., Scholz, M., Thomas, A. M., Valles-Colomer, M., Weingart, G., Zhang, Y., Zolfo, M., Huttenhower, C., Franzosa, E. A., & Segata, N. (2021). Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bobakery 3. *ELife*, *10*. https://doi.org/10.7554/ELIFE.65088

Biocore. (2020). *scikit-bio: A Bioinformatics Library for Data Scientists, Students, and Developers*. GitHub. https://github.com/biocore/scikit-bio

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114. https://doi.org/10.1093/BIOINFORMATICS/BTU170

Bowers, R. M., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T. B. K., Schulz, F., Jarett, J., Rivers, A. R., Eloe-Fadrosh, E. A., Tringe, S. G., Ivanova, N. N., Copeland, A., Clum, A., Becraft, E. D., Malmstrom, R. R., Birren, B., Podar, M., Bork, P., … Woyke, T. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature Biotechnology 2017 35:8*, *35*(8), 725–731. https://doi.org/10.1038/nbt.3893

Brown, C. T., Hug, L. A., Thomas, B. C., Sharon, I., Castelle, C. J., Singh, A., Wilkins, M. J., Wrighton, K. C., Williams, K. H., & Banfield, J. F. (2015). Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature 2015 523:7559*, *523*(7559), 208–211. https://doi.org/10.1038/nature14486

Buchfink, B., Xie, C., & Huson, D. H. (2014). Fast and sensitive protein alignment using DIAMOND. *Nature Methods 2014 12:1*, *12*(1), 59–60. https://doi.org/10.1038/nmeth.3176

Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2020). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*, *36*(6), 1925–1927. https://doi.org/10.1093/BIOINFORMATICS/BTZ848

Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology 2019 20:1*, *20*(1), 1–14. https://doi.org/10.1186/S13059-019-1832-Y

Erb, I., & Notredame, C. (2016). How should we measure proportionality on relative gene expression data? *Theory in Biosciences*, *135*(1–2), 21–36. https://doi.org/10.1007/s12064-015-0220-8

Espinoza, J. L. (2019). *soothsayer: High-level analysis package for (bio-)informatics* (v2019.08). GitHub; GitHub. https://github.com/jolespin/soothsayer

Espinoza, J. L. (2020a). *compositional: Compositional data analysis in Python*. GitHub. https://github.com/jolespin/compositional

Espinoza, J. L. (2020b). *ensemble_networkx: Ensemble networks in Python*. GitHub. https://github.com/jolespin/ensemble_networkx

Espinoza, J. L. (2020c). *hive_networkx: Hive plots in Python*. GitHub. https://github.com/jolespin/hive_networkx

Espinoza, J. L., Dupont, C. L., O'Rourke, A., Beyhan, S., Morales, P., Spoering, A., Meyer, K. J., Chan, A. P., Choi, Y., Nierman, W. C., Lewis, K., & Nelson, K. E. (2021). Predicting antimicrobial mechanism-of-action from transcriptomes: A generalizable explainable artificial intelligence approach. *PLOS Computational Biology*, *17*(3), e1008857. https://doi.org/10.1371/journal.pcbi.1008857

Espinoza, J. L., Harkins, D. M., Torralba, M., Gomez, A., Highlander, S. K., Jones, M. B., Leong, P., Saffery, R., Bockmann, M., Kuelbs, C., Inman, J. M., Hughes, T., Craig, J. M., Nelson, K. E., & Dupont, C. L. (2018). Supragingival Plaque Microbiome Ecology and Functional Potential in the Context of Health and Disease. *MBio*, *9*(6). https://doi.org/10.1128/mBio.01631-18

Fernandes, A. D., Reid, J. N. S., Macklaim, J. M., McMurrough, T. A., Edgell, D. R., & Gloor, G. B. (2014). Unifying the analysis of high-throughput sequencing datasets: Characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, *2*(1), 15. https://doi.org/10.1186/2049-2618-2-15

Franzosa, E. A., McIver, L. J., Rahnavard, G., Thompson, L. R., Schirmer, M., Weingart, G., Lipson, K. S., Knight, R., Caporaso, J. G., Segata, N., & Huttenhower, C. (2018). Species-level functional profiling of metagenomes and metatranscriptomes. *Nature Methods*, *15*(11), 962–968. https://doi.org/10.1038/s41592-018-0176-y

Guo, J., Bolduc, B., Zayed, A. A., Varsani, A., Dominguez-Huerta, G., Delmont, T. O., Pratama, A. A., Gazitúa, M. C., Vik, D., Sullivan, M. B., & Roux, S. (2021). VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome 2021 9:1*, *9*(1), 1–13. https://doi.org/10.1186/S40168-020-00990-Y

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

Hyatt, D., Chen, G.-L., Locascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*, 119. https://doi.org/10.1186/1471-2105-11-119

Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., & Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications 2018 9:1*, *9*(1), 1–8. https://doi.org/10.1038/s41467-018-07641-9

Kang, D. D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., & Wang, Z. (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, *7*(7). https://doi.org/10.7717/PEERJ.7359

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology 2019 37:8*, *37*(8), 907–915. https://doi.org/10.1038/s41587-019-0201-4

Krzywinski, M., Birol, I., Jones, S. J., & Marra, M. A. (2012). Hive plots--rational approach to visualizing networks. *Briefings in Bioinformatics*, *13*(5), 627–644. https://doi.org/10.1093/bib/bbr069

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357. https://doi.org/10.1038/NMETH.1923

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*(7), 923–930. https://doi.org/10.1093/BIOINFORMATICS/BTT656

Lovell, D., Pawlowsky-Glahn, V., Egozcue, J. J., Marguerat, S., & Bähler, J. (2015). Proportionality: A Valid Alternative to Correlation for Relative Data. *PLOS Computational Biology*, *11*(3), e1004075. https://doi.org/10.1371/journal.pcbi.1004075

Mandal, S., Van Treuren, W., White, R. A., Eggesbø, M., Knight, R., & Peddada, S. D. (2015). Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial Ecology in Health & Disease*, *26*(0). https://doi.org/10.3402/mehd.v26.27663

Nabwera, H. M., Espinoza, J. L., Worwui, A., Betts, M., Okoi, C., Sesay, A. K., Bancroft, R., Agbla, S. C., Jarju, S., Bradbury, R. S., Colley, M., Jallow, A. T., Liu, J., Houpt, E. R., Prentice, A. M., Antonio, M., Bernstein, R. M., Dupont, C. L., & Kwambana-Adams, B. A. (2021). Interactions between fecal gut microbiome, enteric pathogens, and energy regulating hormones among acutely malnourished rural Gambian children. *EBioMedicine*, *73*, 103644. https://doi.org/https://doi.org/10.1016/j.ebiom.2021.103644

Nayfach, S. (2021, May). *Recommended cutoffs for analyzing CheckV results?* BitBucket. https://bitbucket.org/berkeleylab/checkv/issues/38/recommended-cutoffs-for-analyzing-checkv

Nayfach, S., Camargo, A. P., Schulz, F., Eloe-Fadrosh, E., Roux, S., & Kyrpides, N. C. (2020). CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nature Biotechnology 2020 39:5*, *39*(5), 578–585. https://doi.org/10.1038/s41587-020-00774-7

Nissen, J. N., Johansen, J., Allesøe, R. L., Sønderby, C. K., Armenteros, J. J. A., Grønbech, C. H., Jensen, L. J., Nielsen, H. B., Petersen, T. N., Winther, O., & Rasmussen, S. (2021). Improved metagenome binning and assembly using deep variational autoencoders. *Nature Biotechnology 2021 39:5*, *39*(5), 555–560. https://doi.org/10.1038/s41587-020-00777-4

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Research*, *27*(5), 824–834. https://doi.org/10.1101/gr.213959.116

Opitz, D., & Maclin, R. (1999). Popular Ensemble Methods: An Empirical Study. *Journal of Artificial Intelligence Research*, *11*, 169–198. https://doi.org/10.1613/JAIR.614

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, *25*(7), 1043–1055. https://doi.org/10.1101/gr.186072.114

Quinn, T. P., Crowley, T. M., & Richardson, M. F. (2018). Benchmarking differential expression analysis tools for RNA-Seq: Normalization-based vs. log-ratio transformation-based methods. *BMC Bioinformatics*, *19*(1). https://doi.org/10.1186/s12859-018-2261-8

Quinn, T. P., & Erb, I. (2020). *Amalgams : data-driven amalgamation for the reference-free dimensionality reduction of zero-laden compositional data*. 1–16.

Quinn, T. P., Richardson, M. F., Lovell, D., & Crowley, T. M. (2017). Propr: An R-package for Identifying Proportionally Abundant Features Using Compositional Data Analysis. *Scientific Reports*, *7*(1), 1–9. https://doi.org/10.1038/s41598-017-16520-0

Ruiz-Perez, C. A., Conrad, R. E., & Konstantinidis, K. T. (2021). *MicrobeAnnotator: a user-friendly, comprehensive functional annotation pipeline for microbial genomes*. *22*(1), 1–16.

Salibian-Barrera, M., & Zamar, R. H. (2002). Bootstrapping Robust Estimates of Regression. *The Annals of Statistics*, *30*(2), 556–582.

Sieber, C. M. K., Probst, A. J., Sharrar, A., Thomas, B. C., Hess, M., Tringe, S. G., & Banfield, J. F. (2018). Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature Microbiology 2018 3:7*, *3*(7), 836–843. https://doi.org/10.1038/s41564-018-0171-1

Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, *9*(1), 1–12. https://doi.org/10.1038/s41598-019-41695-z

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … Vázquez-Baeza, Y. (2020). SciPy

391 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272.
392 https://doi.org/10.1038/s41592-019-0686-2
393 Wu, Y.-W., Simmons, B. A., & Singer, S. W. (2016). MaxBin 2.0: an automated binning algorithm to
394 recover genomes from multiple metagenomic datasets. *Bioinformatics*, *32*(4), 605–607.
395 https://doi.org/10.1093/BIOINFORMATICS/BTV638
396

MICROBIOLOGY

Coral microbiome manipulation elicits metabolic and genetic restructuring to mitigate heat stress and evade mortality

Erika P. Santoro[1], Ricardo M. Borges[2], Josh L. Espinoza[3,4], Marcelo Freire[3,5], Camila S. M. A. Messias[1], Helena D. M. Villela[1], Leandro M. Pereira[1], Caren L. S. Vilela[1], João G. Rosado[1,6], Pedro M. Cardoso[1], Phillipe M. Rosado[1], Juliana M. Assis[1], Gustavo A. S. Duarte[1], Gabriela Perna[6,7], Alexandre S. Rosado[1,8], Andrew Macrae[1], Christopher L. Dupont[3], Karen E. Nelson[3], Michael J. Sweet[9], Christian R. Voolstra[6,7], Raquel S. Peixoto[1,6]*

Beneficial microorganisms for corals (BMCs) ameliorate environmental stress, but whether they can prevent mortality and the underlying host response mechanisms remains elusive. Here, we conducted omics analyses on the coral *Mussismilia hispida* exposed to bleaching conditions in a long-term mesocosm experiment and inoculated with a selected BMC consortium or a saline solution placebo. All corals were affected by heat stress, but the observed "post-heat stress disorder" was mitigated by BMCs, signified by patterns of dimethylsulfoniopropionate degradation, lipid maintenance, and coral host transcriptional reprogramming of cellular restructuration, repair, stress protection, and immune genes, concomitant with a 40% survival rate increase and stable photosynthetic performance by the endosymbiotic algae. This study provides insights into the responses that underlie probiotic host manipulation. We demonstrate that BMCs trigger a dynamic microbiome restructuring process that instigates genetic and metabolic alterations in the coral host that eventually mitigate coral bleaching and mortality.

INTRODUCTION

Coral reefs have been undergoing unprecedented mass coral bleaching events in recent decades, fueled by ocean warming (*1*), heightening the need to devise effective countermeasures to mitigate further declines (*2*, *3*). Increasing sea surface temperatures trigger the disruption of the symbiotic relationship between the coral host and its endosymbiotic algae of the family Symbiodiniaceae (*4*), resulting in the physical whitening of coral colonies known as "bleaching." Photosynthetic products from the endosymbiont algae provide more than 90% of the host's nutritional demands (*5*). Thus, prolonged periods of heat stress and bleaching lead to coral mortality (*6*).

Besides endosymbiotic algae, corals are associated with a suite of other organisms (bacteria, protists, fungi, viruses, etc.), collectively referred to as the coral holobiont or metaorganism (*7*–*10*). In particular, bacteria are assumed to contribute to coral holobiont biology, notably stress tolerance and adaptation to disparate environments (*10*–*15*). The importance of bacteria led to the proposal of the coral probiotic hypothesis (*16*), which states that microbes support coral biology through selection of the most advantageous holobiont configuration in a given environment. This was later refined by the microbiome flexibility hypothesis to include the notion that the potential or propensity for microbiome change differs among host species (*15*). The proposal to use these concepts to select and manipulate specific microbes to aid the stress tolerance and resilience of the coral holobiont was dubbed "beneficial microorganisms for corals" (BMCs) (*10*). Beneficial microorganisms putatively support nitrogen fixation, sulfur cycling, scavenging reactive oxygen species (ROS), and production of antibiotics to thwart pathogens, for example (*10*, *11*, *17*).

The proof of concept that manipulating coral microbes improves coral stress tolerance was recently demonstrated in the first experiments to identify the beneficial nature of a selected BMC consortium in ameliorating coral bleaching (*18*). Nevertheless, exactly "how" these BMCs were associated with functional changes in the host remained unknown. Notably, BMCs do not necessarily need to exert their effect on the coral host directly. Hence, the measured holobiont response does not need to be a perfect reflection of the BMC consortium added. Rather, the BMC consortium may benefit the host indirectly, by means of niche occupation, microbial succession, or the prevention of dysbiosis through pathogen deterrence (*10*, *11*, *18*). Furthermore, although the ability of BMCs to ameliorate coral bleaching has been demonstrated (*18*), it is unknown whether they have the capacity to help corals evade mortality, e.g., through the provisioning of alternate metabolites to compensate for the loss of Symbiodiniaceae.

Despite the diversity of the coral microbiome, which makes it challenging to decipher the contribution of associated microbes to coral holobiont biology, the dynamic nature of the coral microbiome, which can often change markedly—e.g., across sites, species, age, and under stress—further hampers the ability to conduct such studies in the natural environment (*15*, *19*–*21*). For this reason, manipulation of BMCs in controlled experimental setups, such as mesocosms (*22*–*24*), provides an avenue to identify important microbial players and study holobiont responses (and

[1]Institute of Microbiology, Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro, Brazil. [2]Walter Mors Institute of Research on Natural Products, Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro, Brazil. [3]Department of Genomic Medicine and Infectious Diseases, J. Craig Venter Institute, La Jolla, CA, USA. [4]Applied Sciences, Durban University of Technology, Durban, South Africa. [5]Department of Infectious Diseases and Global Health, School of Medicine, University of California San Diego, La Jolla, CA, USA. [6]Red Sea Research Center (RSRC), Division of Biological and Environmental Science and Engineering (BESE), King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia. [7]Department of Biology, University of Konstanz, Konstanz 78457, Germany. [8]Division of Biological and Environmental Science and Engineering (BESE), King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia. [9]Aquatic Research Facility, Environmental Sustainability Research Centre, University of Derby, Derby, UK.
*Corresponding author. Email: raquel.peixoto@kaust.edu.sa

putative underlying mechanisms), while maintaining a quasi-reef environment, to improve and inform the development of biotechnological solutions to promote coral reef resilience.

Here, we used coral mesocosms in combination with multiomics evaluation to assess responses and potentially decipher the mechanisms that underlie the increased stress tolerance and coral mortality evasion, offered by the provisioning of probiotics. In a large-scale effort, fragments of the coral *Mussismilia hispida* were exposed to thermal stress in a 75-day mesocosm experiment and inoculated with either a *M. hispida*–tailored BMC consortium or a saline solution placebo. Coral health (measured via F_v/F_m rates and survivorship) (*25*), microbial activity, and functional responses were assessed through a multiomics approach. Our analysis shows that increased stress tolerance and survivorship of coral holobionts exposed to a BMC consortium coincided with holobiont restructuring and a defined reprogramming of the coral host's gene expression, targeting cellular reconstruction, immune response, and stress protection during a post-heat stress recovery period.

RESULTS

BMC consortium selection, assembly, and experimental setup

Bacterial strains were isolated from a visually healthy colony of *M. hispida*. The BMC consortium was assembled with bacterial strains exhibiting (i) at least one of the beneficial traits detailed below, (ii) the absence of antagonist activity against other selected BMCs, and (iii) no previous record of the species/strain being harmful to humans or other marine life. Beneficial traits included nitrogen fixation (*nif*H), denitrification (*nir*K), dimethylsulfoniopropionate (DMSP) degradation (*dmd*A), ROS scavenging potential (measured through catalase activity), and antagonistic activity against two coral pathogens, *Vibrio coralliilyticus* strain V1 and *Vibrio alginolyticus* V2 (*26*, *27*).

From an initial 133 obtained isolates, the assembled BMC consortium was composed of the following six bacterial strains: *Bacillus lehensis* (M20) positive for *nif*H, *nir*K, and *dmd*A; *Bacillus oshimensis* (M24) positive for *dmd*A; *B. lehensis* (M3) positive for *nif*H and *dmd*A; *Brachybacterium conglomeratum* (M1) positive for catalase and *nif*H; *Planococcus rifietoensis* (CM29) with antagonistic activity against V1; and *Salinivibrio* sp. (F2) with antagonistic activity against V1 and V2 (table S1). The experimental BMC consortium consisted of lag phase–grown bacterial strains collected and resuspended in sterile saline solution (0.85% NaCl) at 1×10^8 cells/ml (for details, see fig. S1). The placebo/control consisted of a sterile saline solution (0.85% NaCl), hereafter referred to as the placebo treatment.

BMCs and placebo were applied every 3 days during a simulated heat stress event (maximum temperature of 30°C) and every 5 days for the remainder of the 75-day mesocosm experiment (Fig. 1A), while a control was run in parallel (26°C). We focused on four time points, T0 at the beginning of the experiment, T1 upon reaching peak temperature in the heat stress (30°C), T2 at the end of peak temperature heat stress (30°C for 10 days), and T3 following a 15-day recovery period at 26°C. Microbiome changes associated with BMC treatment were investigated through 16*S* ribosomal RNA (rRNA) gene metabarcoding (T0, T1, T2, and T3). In addition, patterns and mechanisms underpinning the projected increased stress resilience provided by the BMC treatment were assessed through the evaluation of coral physiology (photosynthetic efficiency of Symbiodiniaceae and visual monitoring of bleaching for T0, T1, T2, and T3) (Fig. 1, B and C), elucidation of metabolic footprints [nuclear magnetic resonance (NMR)/partial least squares-discriminant analysis (PLS-DA) at T0, T1, T2, and T3], and determination of coral transcriptome patterns at the peak of temperature and the end of the experiment (T2 and T3) (see Fig. 1A for experimental design).

Host microbiome shift associated with BMC treatment during heat stress

To confirm the presence of the BMC consortium members in the coral microbiome, the 16*S* rRNA gene sequences of each of the six BMC members was used to query amplicon sequence variants (ASVs) from both BMC- and placebo-treated coral fragments. We identified three of the six strains throughout various time points of the experiment: CM29 *P. rifietoensis* (T1 BMC-treated), M24 *B. oshimensis* (T1 and T2 BMC-treated), and M1 *B. conglomeratum* (T2 BMC-treated) (Fig. 2A).

Parallel to the confirmed microbiome incorporation of three of the BMC strains in T1 (CM29 and M24) and T2 (M1and M24) (Fig. 2A), the overall bacterial community structure of BMC-treated corals was significantly different from placebo-treated corals during the heat stress (T2) [permutational multivariate analysis of variance (PERMANOVA), $P = 0.05$] but became indiscernible during the recovery period (T3) ($P = 0.583$, stress = 0.15) (fig. S2), where BMC strains were also not detected. The most abundant bacterial phyla identified across all coral fragments were Proteobacteria, followed by Bacteroidetes, throughout the course of the experiment (fig. S3). Despite such consistency at higher taxonomic levels, we found variability over time with regard to bacterial taxa association. Although the relatively most abundant genera associated with corals at T0 were consistently *Ruegeria* (11.9%), *Tistlia* (4.6%), and *Candidatus Amoebophilus* (4.2%), we only found *Ruegeria* species to be abundant across BMC-inoculated corals (T1: 15.1%; T2: 13.4%; T3: 17.3%), while in placebo-treated coral fragments, *Paramaledivibacter* spp. were the most prevalent (T1: 21.6%; T2: 13.8%; T3: 3.9%). In addition, ASVs exhibiting significant differences in abundance were identified in BMC-treated coral fragments under thermal stress compared to placebo samples (Fig. 2B). Overall, 13 ASVs were significantly increased [average fold change (FC) = 22.4, $P < 0.01$] in BMC-treated samples in T1, 23 ASVs in T2 (average FC = 21.8, $P < 0.01$), and 18 ASVs in T3 (average FC = 21.7, $P < 0.01$) (Fig. 2B), indicating that BMCs affected the microbiome structure beyond the addition of selected strains. *Ruegeria* was the most prominent genus found in BMC-inoculated corals and was mainly enriched in T2 samples (FC = 15.3) (Fig. 2B). Despite the observed microbiome structural changes, overall community diversity of BMC- and placebo-treated corals remained similar throughout the course of the experiment [based on Shannon, Chao1, and ASV distribution indexes; analysis of variance (ANOVA), $P = 0.8$] (fig. S4).

Coral BMC treatment contributes to increased survivorship and recovery from bleaching after acute thermal stress

We compared photosynthetic efficiency (Fig. 1B) and coral holobiont survival (Fig. 1C) to assess the BMC treatment effect. Most notably, survivorship of corals inoculated with the BMC consortium was substantially higher, with 100% of fragments surviving the heat

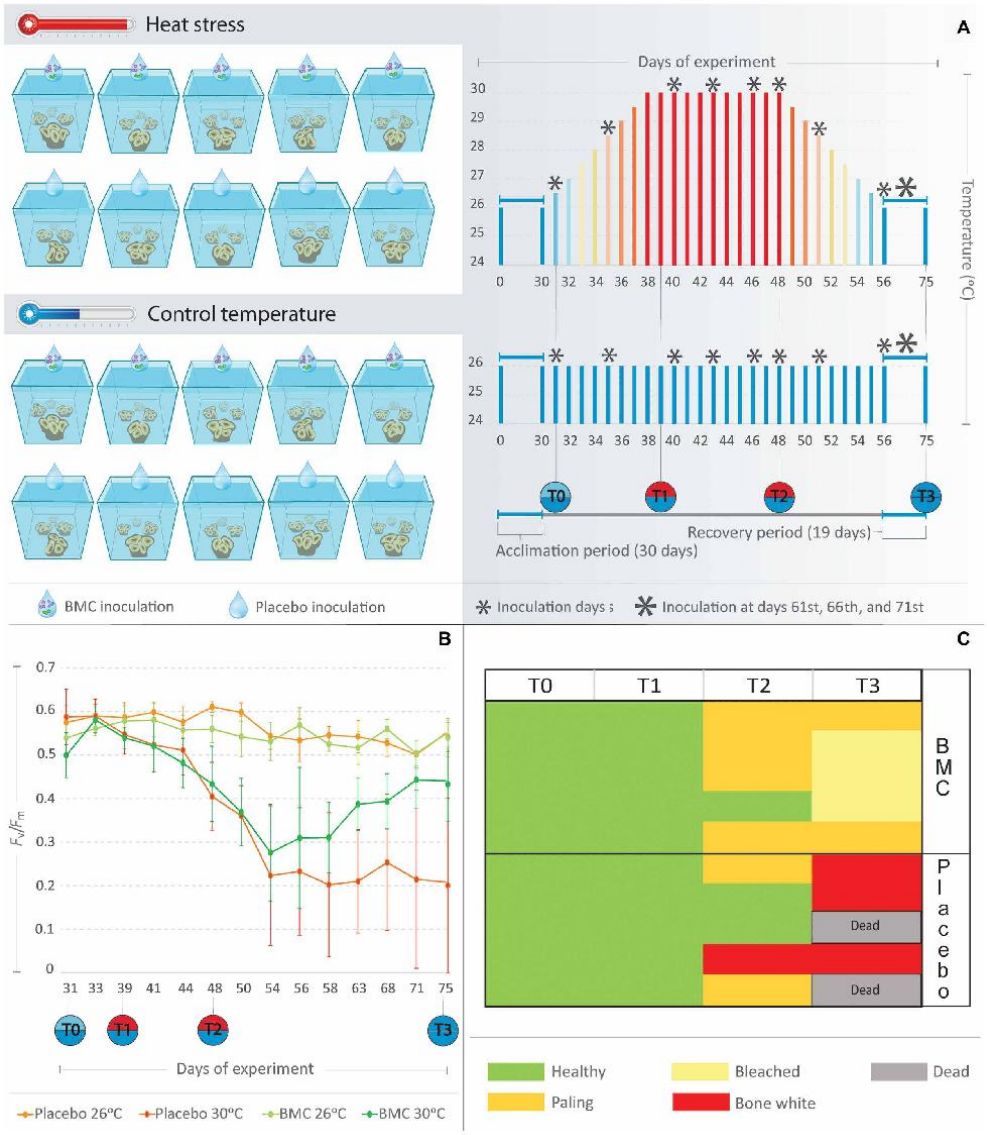


Fig. 1. Long-term heat stress experiment and coral bleaching responses to placebo and BMC inoculation. (**A**) Experimental design and details on temperature, BMC inoculations, and sampling layout. (**B**) Means of photosynthetic efficiency F_v/F_m ratios (*y* axis) from coral fragments treated with BMCs or placebo under heat stress temperature regimes (30°C) and control temperature regimes (26°C) during the mesocosm experiment days (*x* axis). (**C**) Heatmap based on the bleaching score attributed to coral fragments treated with BMCs or placebo in the heat stress experiment.

stress treatment (T3) compared to only 60% of the placebo-treated corals (Fig. 1C). Surviving corals in the placebo-treated regime showed a significant decrease in the F_v/F_m average rates (65% decrease, from T0 to T3; $P < 0.05$) at the end of the experiment compared to the start (Fig. 1B), while photosynthetic efficiencies of BMC-treated corals only decreased at the peak of temperature stress (T2) ($P < 0.05$) and thereafter returned to the initial average during the recovery period (T3) ($P = 0.197$).

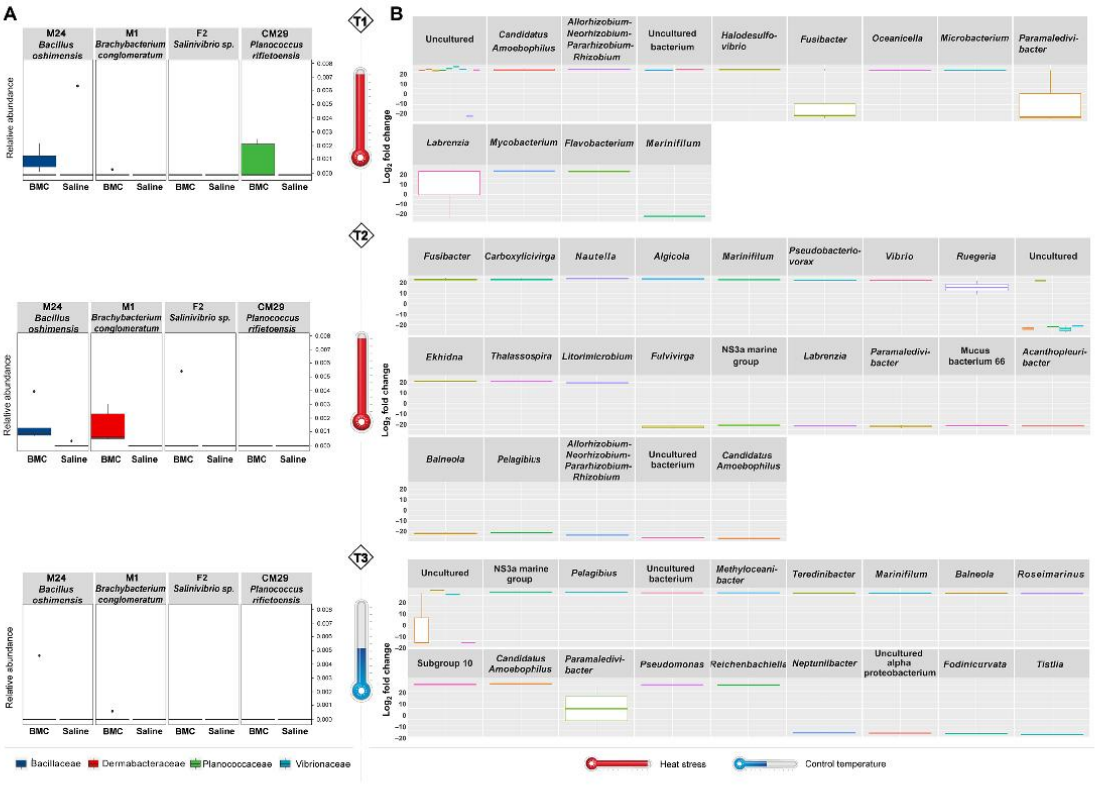


Fig. 2. Effects of BMC treatment on coral bacterial community. (**A**) Relative abundance of BMC consortium members in coral fragments treated with BMCs or placebo and exposed to heat stress (T1, $P = 0.028$; T2, $P = 0.0001$; T3, $P = 0.265$; Kruskal-Wallis), where boxes represent the relative mean abundance and stars represent outliers. (**B**) Boxplot of fold change (FC) of ASVs with differential abundance ($P = 0.01$) in BMC-treated coral fragments compared with placebo-treated fragments at T1, T2, and T3. Bars with the same color scale belong to the same taxonomic family.

Post-heat stress disorder and transcriptional reprogramming of BMC-treated coral holobionts

We were further interested in elucidating the coral host transcriptomic response associated with the observed increase in coral survival after heat stress following BMC treatment. RNA sequencing (RNA-seq) was conducted on samples from BMC- and placebo-treated coral fragments at the peak of heat stress (T2; $n = 20$ samples) and at the end of the experiment (T3; $n = 15$ samples) from the control and heat stress temperatures. Coral genes were assorted into orthogroups to increase confidence in their annotation and, hence, functional inference. We determined a total of 17,755 orthogroups considered for the gene expression analysis (table S2). As expected, we observed pronounced differences in the response to heat stress when comparing placebo-treated corals at 30° and 26°C at T2 (peak of heat stress) [differential expression of 2294 orthogroups with a false discovery rate (FDR) of <0.05 associated with metabolic disorders, apoptosis, autophagy, and response to stress] (table S2). Significant response differences were also observed after a period of recovery (T3), suggesting heat stress carry-on effects in the transcriptional footprint, which we termed post-heat stress disorder (PHSD) with some signs of recovery (2275 orthogroups with an FDR of <0.05 associated with metabolism, cell death, and oxidative stress) (table S2).

Similar to the physiological and metabolic responses, we did not see significant transcriptomic differences between BMC- and placebo-treated coral samples in T2 in the heat stress treatment, suggesting that both "holobiont systems" react similarly in the peak of heat stress, originally observed in yeast and termed environmental stress response (*28*, *29*). Following this, we focused on differentially expressed orthogroups between BMC- and placebo-treated coral samples subsequent to the heat stress at the recovery time point (T3) to elucidate the transcriptomic footprint associated with BMC-induced recovery (fig. S5). BMCs seemed to exert an overall "healing effect," as evidenced by increased recovery and stress attenuation processes in coral gene expression. In this regard, a total of 169 orthogroups were differentially expressed because of BMC inoculation, mainly involved in apoptosis, inflammatory response, cytoskeleton, and membrane reorganization (see blue bars for up-regulation and red bars for down-regulation; Fig. 3 and fig. S5). Most of these

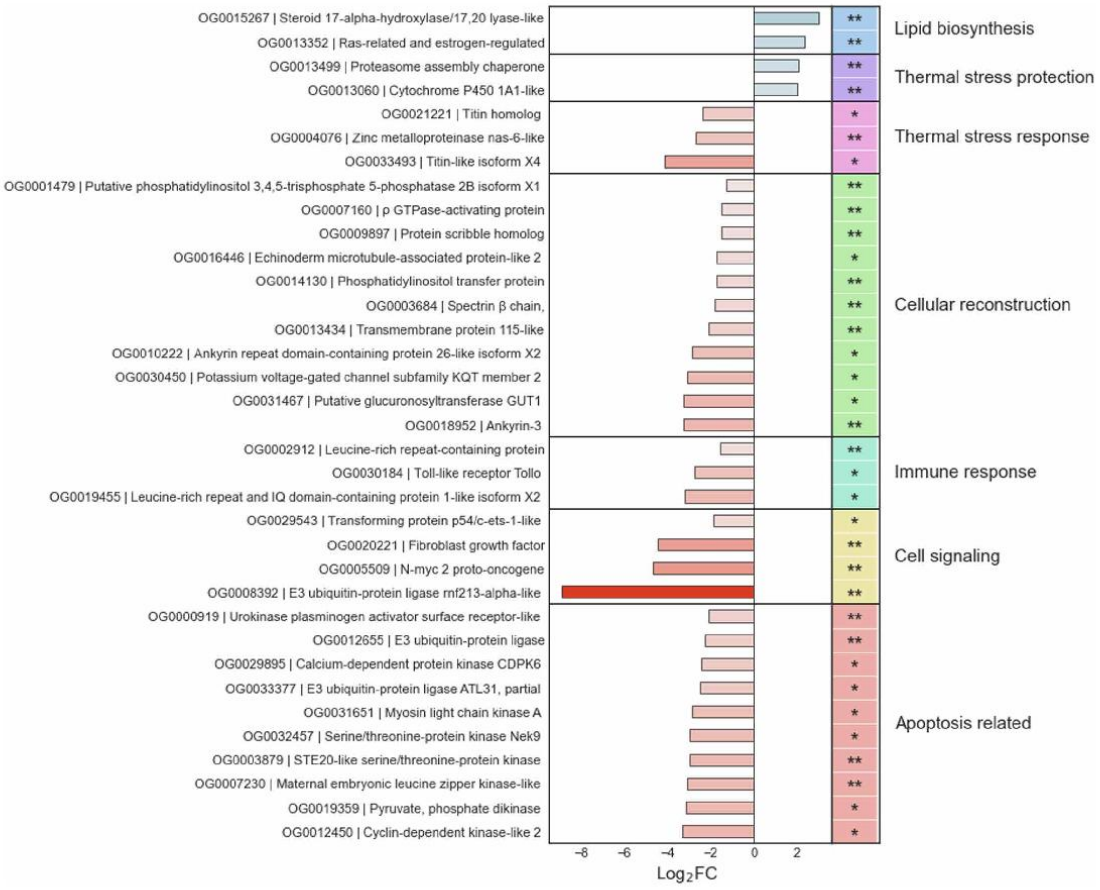


Fig. 3. Coralhost responses to BMC treatment. Main orthogroups with significant (FDR < 0.05) differential expression between BMC- and placebo-treated corals from the end of the heat stress temperature experiment (T3). The respective Kyoto Encyclopedia of Genes and Genomes annotation, their FDR value, and $\log_2$ FC representing up-regulation (positive values and blue bars) or down-regulation (negative values and red bars) in relation to BMC samples are also shown. Orthogroups marked with ** are conserved among *Scleractinia*. Orthogroups marked with * are exclusively from *M. hispida*

orthogroups (142 of 169) were down-regulated in BMC-treated corals in comparison to their placebo-treated counterparts, suggesting attenuation of the PHSD due the probiotic treatment.

Even after the recovery period, the remaining placebo-treated coral survivors were still showing signs of prolonged PHSD, as suggested by the more prominent expression of proteins involved in response to thermal stress, when compared to BMC-treated corals. In addition, orthogroups involved in chromosome condensation (two titin homolog proteins) and DNA methylation (zinc metalloproteinase nas-6-like) were also down-regulated in BMC samples ($\log_2$ FC = −4.12, −2.34, and −2.82, respectively). Overall, the ongoing PHSD observed was significantly mitigated in BMC-treated corals. Numerous orthogroups involved in triggering apoptosis were more highly expressed in placebo-treated corals. Following BMC treatment, we observed down-regulation of seven kinases and one kinase receptor, as well as E3 ubiquitins involved in apoptosis signaling ($\log_2$ FCs from −8.9 to −2.1; see Fig. 3). Further, mitogen-activated protein kinase signaling orthogroups, such as N-myc 2 proto-oncogene ($\log_2$ FC = −4.66) and transforming protein p54/c-ets-1-like ($\log_2$ FC = −1.87) involved in kinase-signaling activation, were down-regulated by BMC treatment. PHSD seemed to also trigger inflammatory and innate immune responses, not only potentially through activity of some kinases but also due to the increased expression of Toll-like receptor, leucine-rich repeat protein and domain. By comparison, these orthogroups were all down-regulated in BMC-treated corals, following heat stress ($\log_2$ FC = −2.73, −1.55, and −3.17, respectively).

In addition, various orthogroups involved in cytoskeleton organization and anchoring were higher expressed in placebo-treated

corals at the end of the recovery period (T3) and, conversely, down-regulated in BMC-treated corals. Orthogroups annotated as echinoderm microtubule-associated protein-like (log_2 FC = −1.7), ρ guanosine triphosphatase (GTPase)–activating protein (log_2 FC = −1.51), putative phosphatidylinositol 3,4,5-trisphosphate 5-phosphatase 2B isoform X1 (log_2 FC = −1.28), and spectrin β chain (log_2 FC = −1.82) were down-regulated in BMC samples. Specifically, the spectrin β chain orthogroup is a complex that is anchored in the cytoplasm via ankyrin proteins, which were down-regulated in BMC-treated corals after the recovery period [ankyrin repeat domain-containing protein 26-like isoform X2 (log_2 FC = −2.89) and ankyrin-3 (log_2 FC = −3.25)]. In addition, orthogroups associated with the synthesis of membrane or secondary cell wall components, such as phosphatidylinositol transfer protein (log_2 FC = −1.28) and glucuronosyltransferase GUT1 (log_2 FC = −3.25), were less expressed in BMC-treated corals, as well as cellular adhesion proteins, represented by the transmembrane protein 115-like (log_2 FC = −2.12), potassium voltage-gated channel subfamily KQT member 2 (log_2 FC = −3.09), and a scribble homolog (log_2 FC = −1.51) orthogroups. Conversely, all these orthogroups associated with the cellular response to cope with the prolonged PHSD were more prominently expressed in placebo-treated corals.

On the other hand, following BMC treatment, we found up-regulated expression of 32 orthogroups (table S2), suggesting that BMC treatment resulted in the induction of thermal stress protection and blockage of PHSD through increased expression of the proteasome assembly chaperone (log_2 FC = 2.06) and cytochrome P450 1A1-like (log_2 FC = 1.99) orthogroups. In addition, the crucial up-regulation of orthogroups associated with biosynthesis of estrogen and steroids (log_2 FC = 3.0 and 2.31), critical components of cell membranes, was also observed in BMC-treated corals.

Significant expression differences were also observed when comparing corals inoculated with BMCs or placebo kept at 26°C 17 days after the beginning of the manipulation of their microbiomes (T2), as represented by 2371 orthogroups with FDR < 0.05 mainly associated with the up-regulation of metabolic pathways (specially biosynthesis of fatty acids, cholesterols, and steroids) and cellular signaling and cycle in BMC-treated corals (table S2). However, no long-term BMC reprogramming took place when no thermal stress was applied, as represented by the lack of differential ortholog expression at T3 (i.e., 44 days after the beginning of the microbial therapy at 26°C).

Together, we found that inoculation with BMCs instigated restructuring of the transcriptional network and cellular homeostasis, up-regulating key orthogroups associated with PHSD mitigation, such as steroids biosynthesis and stress protection proteins, although the general pattern suggested dampening PHSD through down-regulation of stress-related downstream pathways, e.g., apoptosis (Fig. 3). The proximate cause of the transcriptional reprogramming was the BMC treatment that resulted in a restructured host microbiome, which, in turn, suggests a signal cascade from the microbes to the coral host, during the recovery period, corroborating the notion that the holobiont is the functional biological unit.

Metabolic restructuring of BMC-treated coral holobionts after heat stress

We obtained metabolic profiles from the thermal stress experiment using NMR to identify metabolic mechanisms associated with the microbial and genomic restructuring underpinning the increased thermal tolerance of BMC-treated corals. Sample complexity led to strong overlapping ^{1}H resonances, challenging the elucidation of metabolic patterns (Fig. 4A). Nevertheless, the characterized peaks

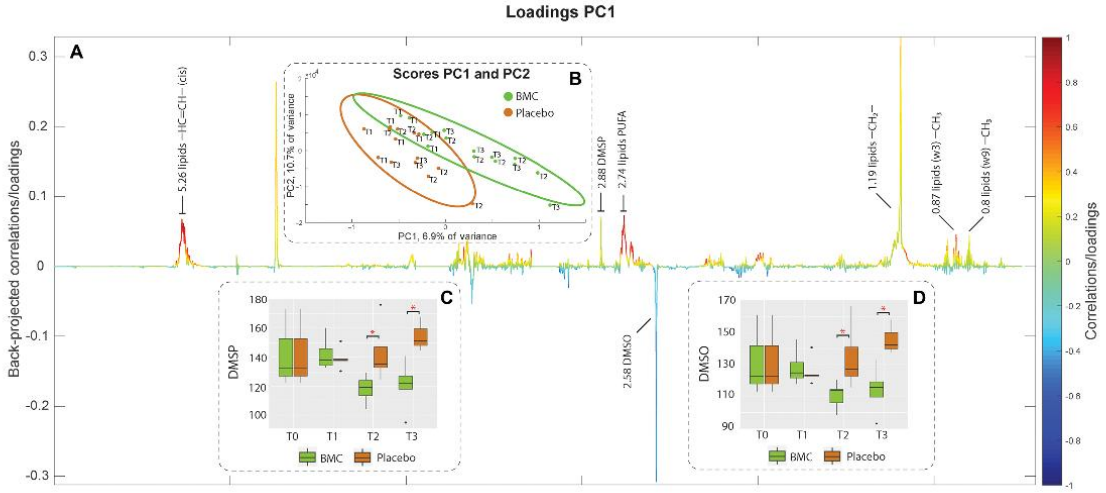


Fig. 4. Metabolic restructuring due to BMC treatment and heat stress. Color-coded loading plot (**A**) (in which colors indicate variation intensity) and score plot with 95% confidence ellipses showing sample clustering by PLS-DA (**B**) from PLS-DA of the ^{1}H NMR dataset comparing the metabolic patterns from coral fragments treated with BMCs and placebo during the thermal stress experiment. Peaks from the loading plot (resonances from annotated as lipids, DMSP, and DMSO) pointing upward are correlated with BMC-treated samples (grouped in the positive quadrant of PC1 (in the score plot), and those pointing downward are correlated with the placebo samples [grouped in the negative quadrant of PC1 (Principal Components 1) in the score plot]. Boxplots are provided to access semiquantitative evaluations of the characteristic DMSP peak at 2.88 ppm (**C**) and DMSO peak at 2.58 ppm (**D**) across sampling time and treatment independently.

at 2.88 parts per million (ppm) (singlet from the S-Methyl groups) for DMSP and at 2.58 ppm (singlet from the S-Methyl groups) for dimethyl sulfoxide (DMSO) were found to be separated and well defined, as well as correlated with different treatments (Fig. 4B). DMSP variation was found to be positively correlated to the BMC-treated corals, and DMSO variation was found to show a negative correlation to the BMC-treated corals in the PLS-DA loading plot (Fig. 4B). To explore the correlations in more detail, comparing metabolic variations overtime (i.e., T1, T2, and T3), the area under the curve representing the direct quantitative ratio for the selected DMSP and DMSO peaks was integrated and represented as boxplots, which indicated significant decreases in the DMSP and DMSO levels observed at T2 and T3 in BMC-treated samples exposed to thermal stress ($P < 0.05$; Fig. 4Cfor DMSP and Fig. 4D for DMSO). We therefore used both DMSP and DMSO as important proxies for the metabolic assessment because of their clear separation from overlapped profiles, as well as their importance in sulfur cycling and microbial structuring (i.e., role of DMSP-related chemotaxis of *V. coralliilyticus* and antimicrobial activity of DMSO). In addition, lipids, despite their predominant presence in every sample, were also positively correlated to the BMC-treated corals as indicated by the PLS-DA loading plot (Fig. 4B). Nevertheless, the strong overlapping signals (at ~5.26, ~2.74, 1.19, and 0.87 ppm), representing positive and negative trends of a unique compound, prevented the possibility of annotating specific compounds. Future efforts should include the analysis of broader molecular spectra by using liquid chromatography–mass spectrometry.

DISCUSSION

The promise of coral probiotics to increase the stress tolerance of corals has been very recently shown (*11*, *18*, *30*), although the effect that BMCs exert on the holobiont or whether BMCs can increase survivability of corals under stress remained elusive. Here, we show that the inoculation of coral fragments with a native BMC consortium instigated holobiont changes at the level of the microbiome, host gene expression, and metabolism, which coincide with an increase in coral survival rates (Fig. 5). Hence, our results provide a first insight into the putative mechanistic underpinnings of how the coral (host) responds to BMC inoculation, although the detailed functional changes that cause the altered phenotype await further elucidation. Our results argue for PHSD recovery improvement of the metaorganism by the BMC consortium, as indicated by changes at the coral host, Symbiodiniaceae, and bacterial compartment level. From the results obtained, a number of key findings emerge that we discuss in the following.

We observed major changes in microbial community structure observed during heat stress (*12*, *31*, *32*) in conjunction with the dynamic microbiome restructuring following the recovery period, indicating that *M. hispida* exhibits microbiome adaptation. Thus, it may fit into the "microbiome conformer" type previously suggested (*14*, *15*) and observed for this coral species regarding other impacts (*33*–*35*). Following this notion, the level of microbiome flexibility may be considered as a factor to identify corals with high(er) manipulative potential. Corals that naturally alter their microbial composition and potentially uptake microbes from the environment are more likely to "accept" inoculants (*14*, *15*, *36*). Notably, shifts in metaorganism microbial composition are, potentially, rapid and versatile means of adaptation to environmental change (*12*–*15*).

It is important to consider that the host's ability to take up microorganisms from the environment is hypothesized to increase when under stress, a conclusion based on the finding that many host microbiomes appear less ordered when stressed (*14*, *21*, *37*). Inoculation with high numbers of different BMC cells (i.e., a consortium) may therefore ensure (and improve) uptake of at least some microorganisms exhibiting beneficial characteristics, which may, at the same time, preclude colonization by pathogens considering that "space is limited." The use of bacterial consortia provides a combination of beneficial mechanisms to increase stress tolerance, even if not all members of the BMC successfully associate with the coral holobiont (*18*, *31*, *38*–*40*). Here, we show that the use of a bacterial consortium resulted in incorporation of some of the selected BMCs, which were found in the microbiome of BMC-treated corals during the thermal stress, i.e., at T1 and T2 (see Fig. 2A). Notably, members of the BMC consortium were not detected after the 15-day period (T3; i.e., recovery). This suggests three things: first, a dynamic restructuring of the microbiome can happen on a relatively small time scale (*12*, *14*, *41*); second, incorporation of BMCs might be facilitated under stress (in this experiment, during the peak of heat stress) because coral defense is compromised or selection for beneficial microbes is supported; and third, it is currently unclear how long the beneficial effect of BMCs is lasting. From our results, it appears that BMC members colonized coral fragments during stress and instigated significant changes in the coral holobiont but reverted to the original microbiome structure after ceasing (or the absence) of stress [sensu Ziegler *et al.* (*14*) who used the term "microbiome recovery"]. Accordingly, the duration of the presence of the stressor might determine the longevity of the BMC effect, which suggests that repeated addition of BMCs might be needed to ensure a long-lasting effect under natural conditions (*11*).

The early and detectable incorporation of some of the BMC consortium members into the coral microbiome and the subsequent microbial restructuring were correlated with significant improvements in coral recovery after thermal stress, as most convincingly demonstrated by mortality evasion. Heat stress–driven mortality and/or decrease in F_v/F_m rates observed in fragments that were not treated with BMCs suggest damage to the temperature-related photosystem II electron transport of the Symbiodiniaceae through chronic photoinhibition (*42*), which ultimately leads to a breakdown in symbiosis and results in loss/expelling of the Symbiodiniaceae, i.e., bleaching (*43*). Notably, bleaching is a symptomatic phenotype, i.e., loss of Symbiodiniaceae can occur through multiple processes, including host cellular apoptosis (*44*) or necrosis, and eventually death from starvation (*6*, *45*), which was corroborated by the up-regulation of different kinases directly involved in triggering apoptosis in placebo-treated corals (*46*, *47*). Our transcriptome results indicate that BMCs did not buffer the immediate heat stress response in *M. hispida* but exerted its effect during recovery, supported by the gene expression patterns and coral physiology. Most notably, we observed low F_v/F_m rates for both BMC and placebo treatments at T2 (during heat stress), but only the BMC treatment promoted recovery at T3, as indicated by the "return-to-normal" F_v/F_m rates. Our interpretation is that BMCs exert their effect through mitigation of the effects from what we term PHSD. The molecular evidence for this condition includes not only apoptosis activity, which may be triggering inflammatory responses, but also membrane and cellular reconstruction due to tissue loss caused by recent-past heat stress. In this regard, the remaining placebo-treated

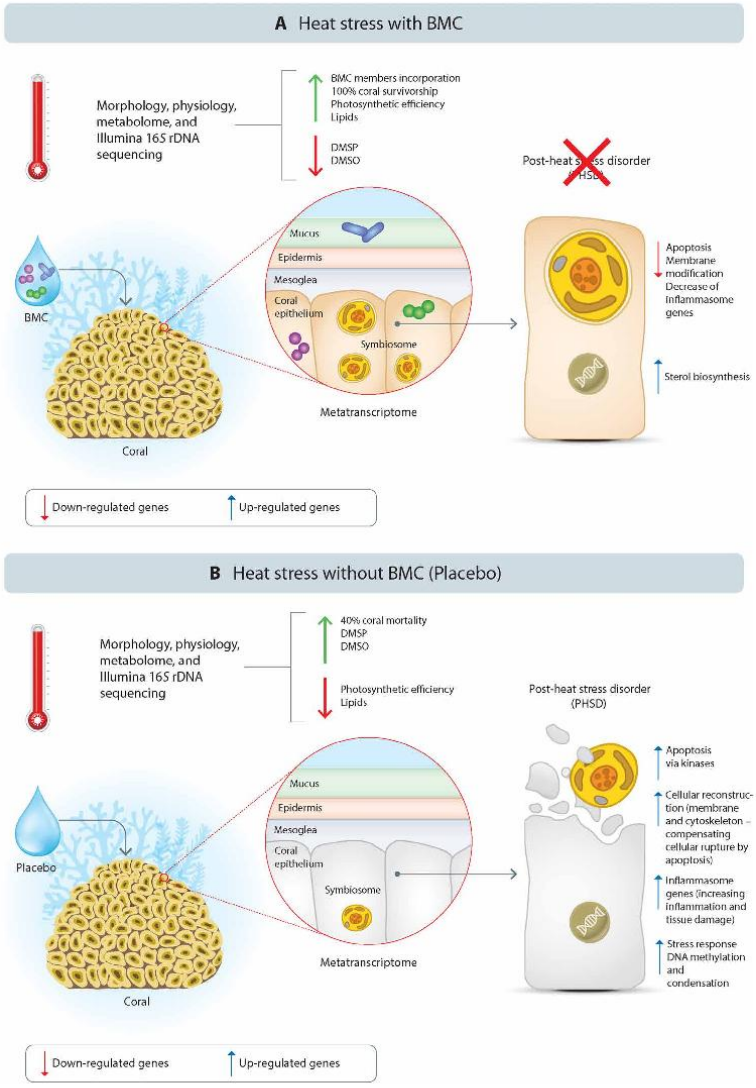


Fig. 5. Probiotics-mediated mitigation of coral PHSD. Summary of the overall differential recovery mechanisms observed at the end of the 75-day mesocosm experiment, comparing the process in BMC-treated (**A**) and placebo-treated (**B**) *M. hispida* fragments.

surviving corals seemed to be still struggling from the effects of recent heat stress, even 19 days after the end of the heat stress period, while all BMC-treated corals seem to have recovered.

As an analogy to posttraumatic stress disorder (*48*), coral PHSD is characterized by the contrast of the coral response and its attempts to recover from a heat stress event while still fading due to the cellular, immune, and metabolic consequences of such stress. The significant up-regulation of numerous kinases and receptors, as well as signaling molecules, by the remaining placebo-treated survivor corals at T3 suggests ongoing apoptosis (*47*). In addition, the oxidative stress increased by thermal photodamage to the photosynthetic apparatus of Symbiodiniaceae might be further contributing to trigger inflammatory responses (*49*). Previous studies have also found expression of immune-related and apoptosis genes in corals

affected by heat stress for extended periods of time (*50*, *51*), suggesting a persistent bleaching effect on the coral transcriptome of susceptible corals (*52*). We hypothesize that down-regulation of orthogroups involved in apoptosis and the concomitant up-regulation of thermal stress protection proteins, such as chaperones, promoted by BMC inoculation, protected corals from tissue damage and Symbiodiniaceae loss, with consequences for coral survival. Such prominent recovery promoted by coral probiotics indicates that if the selection of BMCs based on the hypothetical framework proposed by Peixoto *et al.* (*10*, *11*) already provides measurable benefit and survivorship improvement, more careful selection of BMCs could result in even larger improvements. It is worthwhile and interesting to also highlight that such reprogramming was also observed when no stress was applied (at 26°C), but only at an early stage. While it seems that inoculation of BMCs rapidly trigger change of response norms from the host, long-term BMC reprogramming and exerted effects are only manifested under and subsequent to stress.

It is tempting to speculate that the increased host survivorship observed in this study is a direct consequence of the transcriptomic changes discussed above, which arguably will result in altered metabolic profiles. For instance, the observed changes in the metabolomic profile of corals treated with BMC supports the hypothesis that the selected microbes play a direct role in increasing coral stress tolerance as evidenced by correspondence between selected traits of BMC bacteria (i.e., DMSP degradation) and observed metabolic changes. Shifts in BMC-treated metabolomic profiles were signified by a decrease in the DMSP concentration and lipidic reservoir maintenance. This connects directly to the presence of M24 in the 16*S* rDNA data: M24 was found exclusively in BMC-treated samples at T1 and T2, indicating its incorporation into the coral microbiome, and was selected because of its ability to degrade DMSP (table S1). Notably, DMSP is mostly produced by algae (Symbiodiniaceae), and its degradation generates antimicrobial compounds, helping to control pathogens (*53*–*55*). Peixoto *et al.* (*10*, *11*) suggested that this is a desirable BMC trait. In parallel to the DMSP degradation as one of the direct mechanisms provided by the BMC consortium to ameliorate heat stress, the BMC treatment may have also indirectly influenced DMSP metabolism, through the enrichment of bacteria able to assimilate DMSP, such as *Ruegeria* (*56*), the most abundant genus found in BMC-treated coral samples during the course of the experiment. This genus has been previously observed to inhibit and control the growth or pathogenicity of *V. coralliilyticus* (*57*). These observed traits may have triggered the molecular responses observed by the host. These results highlight the importance of microbiome restructuring to coral resilience (*12*, *15*) and the additional potential role of the *M. hispida* BMC consortium in modulating the microbial colonization and succession of inoculated coral fragments. This parallel colonization/succession/enrichment of beneficial microbes has also been observed in other hosts, including humans, as a result of the use of pre- or probiotics (*58*, *59*).

The increasing frequency and severity of ocean warming events has caused coral die-offs worldwide in the last few years (*1*, *60*–*62*). The development and better understanding of novel interventions to mitigate large-scale coral mortality is one of the climate priorities for the coming decades (*63*, *64*). The results of this study provide three completely novel insights that can aid the development of tools to promote human-accelerated environmental adaptation of corals: (i) BMC treatment and heat stress are both necessary conditions to trigger a long-term BMC thermal protection effect, whereas neither on its own is sufficient; (ii) the BMC thermal protection effect manifests after the heat stress and affects recovery; and (iii) such BMC-promoted protection mitigates coral PHSD, preventing mortality. Our results support the potential of microbiome restructuring to aid in the environmental adaptation of the coral metaorganism to global change (*15*) and identify a suite of microbial-mediated host responses underlying coral survival and recovery to thermal bleaching provided through BMCs. This is most prominently highlighted by the marked increase of 40% in coral survival rates following thermal stress and prior BMC treatment. This was accompanied by overall shifts in the coral microbiome that suggest a dynamic restructuring of the microbiome, due partially to the incorporation of BMC members and the relative increase of other bacteria. We further show that such microbiome restructuring directly affects the host, exerting beneficial effects and PHSD mitigation, as evidenced by transcriptional reprogramming (i.e., down-regulating apoptosis and inflammatory triggering molecules and up-regulating thermal stress protection proteins). In this light, our results reinforce the promise and potential of coral probiotics as an effective tool to rehabilitate coral reefs, particularly because the ability to "recover" is what eventually makes the difference in the real world, i.e., not only the difference in responding to heat stress but also in surviving the heat stress. In this regard, our data also suggest that prophylactic inoculation of BMCs, a few weeks before thermal events, can be advantageous for corals to sustain heat stress as it supposedly allows them to more rapidly and readily recover from thermal stress.

MATERIALS AND METHODS

Ethics approval and consent to participate

Permission for sampling was obtained from the System of Authorization and Information on Biodiversity. The microbial survey permits were obtained from CNPq (National Council for Scientific and Technological Development, Brazil) and SISGEN (National System for the Management of Genetic Heritage and Associated Traditional Knowledge) (number A620FE5).

Sampling procedures

M. hispida colonies were collected by SCUBA diving at the Coroa Vermelha reefs, Santa Cruz de Cabrália County, Bahia, Brazil. Coral colonies were collected at three sites along the reef: site A (16°20′57.99″ S; 038°58′45.00″ W), site B (16°20′39.30″ S; 038°58′38.10″ W), and site C (16°22′02.20″ S; 039°0′15.63″ W), at depths between 1.5 and 10 m on 26 to 29 January 2017. Corals were transported in sterile plastic bags and then packed in Styrofoam boxes containing 800 g of ice and were sent by air cargo to Rio de Janeiro. Upon arrival at the research station, around 13 hours after sampling, coral colonies were transferred to 1500-liter tanks with constant sea water flow and air bubbling for a 3-day preliminary acclimatization period. After that, coral colonies were fragmented using a diamond-based band saw (Gryphon Corp., CA, USA) in ~7-cm fragments with at least three polyps each, sawn in the coenosarc, and placed in the experimental system for acclimatization and healing. About 4 days after sawing, the coral fragments showed the first signs of healing and were kept in acclimation conditions (26°C) until all fragments reached F_v/F_m rates of around 0.6.

Isolation of bacterial strains from coral

Three previously tagged colonies (5 to 15 cm) of the thermally resistant coral *M. hispida* collected from Maraú, Bahia, Brazil (13°56′10.9″ S; 38°55′38.71″ W) were used as a source to isolate BMCs. Two different approaches were used for bacterial isolation. First, 0.5 g of each coral macerate was resuspended in 45 ml of sterile saline solution (0.85% NaCl) and then shaken for 16 hours. After incubation, triplicate subsamples (100 µl) of 10^{-3}, 10^{-4}, and 10^{-5} dilutions were inoculated into petri dishes containing 20 ml of marine agar medium (Marine Agar Zobell 2216, HiMedia Laboratories, Mumbai, India), diluted marine agar medium (Marine Agar Medium 2× diluted with 2.5% NaCl and agar adjusted), 2.5% NaCl Luria-Bertani medium (10 g of tryptone, 5 g of yeast extract, 25 g of NaCl, and 15 g of agar to 1000 ml of distilled water) or marine water medium (1000 ml of sea water and 13 g of agar). In addition, coral fragments of ~0.5 mm were placed directly onto dishes with these culture media. All the plates were incubated at 26°C for 48 hours. A total of 133 bacterial colonies were isolated, based on bacterial colony morphology, with 67 derived from macerated slurries and 52 derived from mini fragments. Each different morphological colony was stored in an ultra-freezer with a final concentration of 20% glycerol and removed when necessary for functional screening.

Functional screening for probiotic and bacterial 16*S* rRNA gene sequencing

Each morphologically different bacterial isolate was screened for beneficial traits for corals, as proposed by Peixoto *et al.* (*10*). Sixty-seven morphologically distinct bacterial strains were recovered from macerated slurries and 52 from microfragments of the coral placed directly onto the agar medium. The isolates were then screened for beneficial traits, as previously outlined by Peixoto *et al.* (*9*), and tested via a proof-of-concept study (*18*). Antagonistic activity against *V. coralliilyticus* YB strain (DSM19607) (V1) and *V. alginolyticus* (BAA450) (V2) was tested by the agar diffusion method (*65*). First, 20 µl of each bacterial strain was spot-inoculated onto 2.5% NaCl LB medium, placing three spots for each strain (representing replicates). The plates were incubated at 26°C for as long as necessary for the strain to grow. The strains were inactivated by chloroform volatilization, followed by pouring 3 ml of semisolid 2.5% NaCl LB medium (0.7% agar) containing the *Vibrio* indicators over the inactivated spots. These plates were then incubated at 28°C for 16 hours, and the antagonistic activity was indicated by inhibition halos around or no detection of *Vibrio* growth over the colony spot. The same procedure was repeated for both V1 and V2 in separate plates. Among the remaining candidates, one strain, identified as *P. rifietoensis* (CM29), was an antagonist against *V. coralliilyticus* YB (DSM19607) (V1), while another strain, identified as *Salinivibrio* sp. (F2), showed antagonistic activity against both V1 and *V. alginolyticus* (BAA450) (V2). The strains were screened for ROS scavenger enzyme activity, based on qualitative (production or no production) and quantitative (bubble amount) catalase production when 50 ml of their liquid culture was mixed with 50 µl of 3% (v/v) hydrogen peroxide.

Nitrogen-cycling genes, as nitrogenase subunits (*nif*H) and denitrification (*nir*K), as well as the DMSP degradation (*dmd*A) gene were screened by polymerase chain reaction (PCR) from the genomic DNA samples (for additional information about primer and PCR cycling, see table S3). From the initial 133 isolated strains, 33 (25%) isolates demonstrated high catalase activity and positive results for the amplification (PCR) of the genes *nif*H (12 strains, i.e., 9%), *nir*K (5 strains, 11%), and *dmd*A (11 strains, 8%). Almost half of the isolates (49% or 65 of 133) were identified as belonging to the genus *Vibrio* and were excluded from the following steps, considering that they are regularly postulated to be coral pathogens (*66*, *67*). A total of 38 strains positive for at least one screened trait described above had their nearly full-length 16*S* rRNA gene PCR-amplified and sequenced (table S3). The sequencing electropherograms were processed using the Ribosomal Database Project II (*68*) to remove low-quality bases. Sequences of each isolate were assembled into contigs using Bioedit 7.0.5.3 (*69*). The bacterial 16*S* rRNA gene sequences were aligned with sequences from the National Center for Biotechnology Information (NCBI) database (*70*). All sequences were deposited in the NCBI database under an individual accession number, given below (see table S4). Bacteria strains identified as potential human or marine pathogens as well as those with antagonistic activity against any member of the selected BMC consortium (assessed by the agar diffusion method cited above) were excluded.

Probiotic preparation

A total of six bacterial strains—M20 *B. lehensis* (NCBI access number MK308622), M24 *B. oshimensis* (MK308624), M3 *B. lehensis* (MK308617), M1 *B. conglomeratum* (MK308603), CM29 *P. rifietoensis* (MK308593), and F2 *Salinivibrio* sp. (MK308616)—were selected to compose the *M. hispida* BMC consortium, based on the beneficial traits cited above and described in table S1. The probiotic consortium suspension contained a total of 10^8 cells/ml. The cell number of each individual BMC strain was estimated by an optical density spectrophotometer [optical density at 600 nm (OD_{600})] [UV-1800 spectrophotometer (Agilent Cary 60, Agilent Technologies)], measurements for cultures grown at 26°C in 100 ml of LB medium for 8, 16, 22, 30, 42, 48, and 54 hours and correlated directly with the number of colony-forming units (CFUs) of each strain at each time point. The CFUs were assessed by subsampling (100 µl) each serial dilution of each strain at each time point, plating on LB agar medium, and incubating under the same conditions. The results were normalized to 1 ml of medium to estimate the cell number at the sampling points (fig. S1). As the probiotic consortium is composed of a diverse combination of bacteria, each strain was collected proportionally at the peak of its growth curve to compose a consortium with a final concentration of 10^8 cells/ml. The cultures were centrifuged at 5000*g* for 2 min, and the cell pellets washed three times with saline solution (0.85% NaCl), followed by centrifugation and resuspension in 50 ml of saline solution.

Mesocosm experimental design

The experimental mesocosm used for this experiment consisted of two water baths (100 cm by 50 cm by 10 cm) per temperature (total of four water baths), where five individual aquariums (each with 1.3 liters of capacity; 15 cm by 11 cm by 12 cm) from each treatment were randomly distributed in the mesocosm. Each completely individualized aquarium was an independent true biological replica, with its own individual sump (8.7 liters) and circulation pump; the sump and aquarium assembly together contained a total of 10 liters of seawater. The water flow between the sump and the aquariums was driven by a water pump (Mini A, Sarlo Better, São Caetano do Sul, Brazil) at a flow rate of 250 ml min^{-1}, providing a 10-fold recirculation of the experimental aquarium volume per hour. Every 2 days, 10% of the sump water was changed and the salinity adjusted

to 34 practical salinity unit with deionized water if necessary. The aquariums were supplied with natural seawater from the Marine Aquarium of Rio de Janeiro research station where the experiment was performed. The replicates received individual continuous air-bubbling circulation through air pumps (HG-370, Sunsun) connected to silicone air hoses and flow controllers. The water in the water bath was homogenized by two aquarium pumps (SB 1000A, Sarlo Better) to maintain homogeneous temperatures, and there was no water exchange between the aquariums and the water baths. Thermostat controls MT-518ri (Full Gauge, Canoas, Brazil) measured and controlled the temperature of each water bath, activating the cooling system or heaters as needed. The water baths were connected to a 1000-liter freshwater reservoir at 18°C to provide cooling water through water pumps (Better 2000, Sarlo Better) when the temperature-controlled thermostat activated the corresponding pump. The heating system consisted of two 100-W heaters (Atman, China) in the water bath. Physical-chemical parameters of the water, including pH, salinity, and dissolved oxygen (OD), were measured on the sampling days, using a multiparameter probe (Model HI 9,828, Hanna Instruments, Barueri, São Paulo). The experiments followed artificial day/night cycles (12 hours/12 hours) with 150 µmol of photons m^{-2} s^{-1} from 06:00 to 10:00 and from 14:00 to 18:00 hours, and 250 µmol of photons m^{-2} s^{-1} from 10:00 to 14:00 hours, modulated with light dimmers and a shade cloth. Each replicate individual aquarium had its own lighting system, consisting of six 3W of blue-light and three 3W white-light light-emitting diodes (LEDs), each controlled by a potentiometer. Four coral fragments (~7 cm) of *M. hispida* were placed randomly in each aquarium, and a single fragment was randomly used as a sampling unit for each treatment and sampling time.

Mesocosm experiment

A total of 80 coral fragments of *M. hispida* were exposed to two temperature regimes, 30°C (heat stress temperature regime) and 26°C (control temperature regime), and two treatments, placebo or BMCs. A total of four coral fragments (~7 cm) were placed randomly in each aquarium, consisting of a completely independent replica, and each treatment used five of these aquaria. One fragment was randomly used as a sampling unit for each treatment and sampling time. All coral fragments were first maintained under the same conditions at 26°C for 30 days to allow them to heal and acclimate to the experimental conditions. For the heat stress temperature experiment, the temperature was increased, from day 0 to day 8 by 0.5°C per day up to 30°C, which was maintained for 10 days. Then, the temperature was decreased to 26° by 1°C per day, followed by 23 days of recovery. All control experiment aquariums were maintained at 26°C during the 75 experimental days. Sampling points were before heat stress (T0), at the peak of temperature (T1), at the last day of high temperature (T2), and after the recovery period (T3). Samples from the control temperature experiment were also taken in parallel at the same time points. The placebo and BMCs were inoculated on the first day of the experiment and every 5 days thereafter; during the 10 days at the temperature peak, inoculations were performed every 3 days. A detailed schematic view of the experimental design is shown in Fig. 1A. Inoculations were performed by removing the coral fragments from the aquarium and placing them in a sterile petri dish to inoculate 1 ml of the respective treatment above the fragments. After the inoculation, the fragments were immediately returned to their respective aquariums, and the individual petri dishes were rinsed into the aquarium water. Raw data generated in this work are available in the NCBI Sequence Read Archive under the BioProject accession number PRJNA649484.

Assessment of coral health and microbiome

Coral health was assessed during the experiment using different proxies, including visual monitoring of bleaching and algal photosynthetic parameters. The coral visual response was assessed by color score based on the tissue appearance: (i) white (>80% of colony white, with no visible pigmentation), (ii) pale (>10% colony affected by pigment loss), or (iii) fully pigmented (<10% colony with pale coloration). Coral mortality was scored as 0. Each replicate was photographed at each sampling time, with a Canon T3i digital camera, under the same conditions, and the color was scored on the basis of the photographic assessment.

The photochemical efficiency of the Symbiodiniaceae was assessed using pulse amplitude–modulated (PAM) fluorometry. We used a submersible diving-PAM system (Walz GmbH, Effeltrich, Germany) fitted with a red-emitting diode (LED; peak at 650 nm). To avoid nonphotochemical processes of dissipation of PSII excitation energy, measurements were taken after sunset, after at least 30 min of darkness, to ensure full photochemical dissipation of the reaction centers. The maximum quantum yield of PSII photochemistry was determined as F_v/F_m. The diving PAM was configured as follows: measuring light intensity = 5; saturation pulse intensity = 8; saturation pulse width = 0.8; gain = 2; and damping = 2. The same coral fragment from each replicate (= 4 fragments) was used to measure chlorophyll fluorescence at different sampling times during the experiment. The statistical significance of the results was analyzed in Paleontological Statistics software (PAST3).

Assessment of coral microbiome through 16*S* rRNA gene amplicon sequencing

The coral microbiome was assessed by 16*S* rRNA gene amplicon sequencing analysis. Samples of the mucus layer, tissue, and skeleton of the coral were collected with sterile clippers at the sampling time. Samples from each sampling time point were macerated with a mortar and pestle under dry conditions. Total DNA was extracted from 0.5 g of the macerated mucus, tissue, or skeleton using the PowerBiofilm DNA Isolation Kit (MO BIO Laboratories Inc.), following the manufacturer's instructions. The DNA concentration was determined using the Qubit 2.0 Fluorometer High Sensitivity DNA Kit (Invitrogen, USA).

To amplify the hypervariable regions V5 and V6 of the bacterial 16*S* rRNA gene, the primers 784 forward (5′-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGGATTAGATACCCTGGTA-3′) and 1061 reverse (5′-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCRRCACGAGCTGACG AC-3′) (*71*) were used (Illumina adapter sequences underlined). Triplicate PCRs (using 1 µl of input DNA) were performed with the QIAGEN Multiplex PCR kit, with a final primer concentration of 0.3 µM in a final reaction volume of 10 µl. In addition to samples, null template PCRs were run (no template DNA input) to account for putative kit contaminants. Thermal cycler conditions were as follows: initial denaturation at 95°C for 15 min, 27 cycles of 95°C for 30 s, 55°C for 90 s, and 72°C for 30 s, followed by a final extension at 72°C for 10 min. Then, 5 µl of each PCR product was run on a 1% agarose gel to confirm successful amplification. Triplicate PCRs for each sample were pooled and samples cleaned using the ExoProStar 1-Step (GE Healthcare, UK).

Samples were indexed using the Nextera XT Index Kit v2 (dual indices and Illumina sequencing adaptors added). Successful addition of indexes was confirmed by comparing the length of the initial PCR product to the corresponding indexed sample on a 1% agarose gel. Samples were cleaned and normalized using the SequalPrep Normalization Plate Kit (Invitrogen, Carlsbad, CA, USA). The samples were then pooled in an Eppendorf tube (4 µl per sample) and concentrated using the CentriVap Benchtop Vacuum Concentrator (Labnoco, USA). The quality of the library was assessed using the Agilent High Sensitivity DNA Kit in the Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) and quantified using Qubit (Qubit dsDNA High Sensitivity Assay Kit, Invitrogen). Library sequencing was performed at 5 pM with 20% phiX on the Illumina MiSeq Illumina platform at the King Abdullah University of Science and Technology (KAUST) Bioscience Core Lab at 2 × 301 bp paired-end V3 chemistry, according to the manufacturer's specifications.

Coral microbiome data analysis

The microbiota associated with *M. hispida* fragments was investigated by sequencing the V5 to V6 variable region of the 16*S* rRNA gene. A library of 3,501,072 good-quality reads with a mean length of 283.64 bp was generated. Demultiplexed raw sequences were imported into QIIME2 2019.4 for analysis. Sequences were merged, denoised, dereplicated, clustered, and trimmed using the DADA2 ("dada2 denoise-paired") plugin with the following parameters: -p-trim-left-f 5 --p-trim-left-r 5 --p-trunc-len-f 250 --p-trunc-len-r 250, and 4775 ASVs were obtained. The ASVs were classified taxonomically using the naive Bayes machine-learning classifier (*72*) with the q2-feature-classifier parameter, using the SILVA132 (*73*) trained classifier clustered at 99% identity as the reference database. A rooted phylogenetic tree was created for downstream analyses, using the programs MAFFT2 (*74*) and FastTree with CAT-like rate approximation category through Q2-alignment and Q2-phylogeny plugins. The microbial QIIME2 output (qza files) were imported to R programming language version 3.6.0 with the function qiime2R and analyzed with the Phyloseq (*75*), parsed with the dplyr package (*76*), and the barplots, boxplots, and statistical test (Kruskal-Wallis and ANOVA) were generated with ggplot2 (*77*). The data were tested for differential abundance using DESeq2 (*78*) on a model of negative binomial distribution (NEB); a Wald test with parametric fitting of dispersions to the mean intensity was used for differential abundance estimation, using a cutoff value of $P = 0.01$. For nonmetric multidimensional scaling analyses, the data were $\log_2$-transformed ($\log x + 1$), and ordination was performed with the Bray-Curtis distance matrix until a solution was reached (i.e., stress). The results were plotted with function plot_ordination using the samples, treatments, or environment metadata for features ordination. The significance of the results was evaluated with PERMANOVA, using 999 random permutation tests with pseudo *F*-ratios through the Adonis function of the Vegan package (*79*) in R. The community structure of the microbiome (represented by diversity and richness measures) was evaluated (from 0 to 3000 reads) using classic ecological indexes of α diversity (observed ASVs, Chao1, and Shannon) on the rarefaction curve plateau using Phyloseq package.

Gene expression analysis

The coral host gene expression response from T2 and T3 samples were explored to elucidate patterns and genes associated with the physiological differences observed between the BMC and placebo treatments. Total RNA was extracted from both the BMC and placebo treatment samples from the heat stress temperature (30°C) and control temperature (26°C) experiment for sequence analysis, from T0 (five representative replicas), T2 (20 samples, five replicas of each treatment and temperature), and T3 (15 samples, five and three replicas from BMC- and placebo-treated corals of heat stress temperature; four and three replicas from BMC- and placebo-treated corals of control temperature; some of the samples from control temperature had not enough material for RNA extraction). RNA-seq was performed using the Illumina HiSeq 6000 platform (Illumina Inc., San Diego, CA, USA). Sequence reads were quality-controlled using KneadData v0.7.4 with the GRCh38.p13 human genome as a reference for potential decontamination, yielding between 14,372,271 and 74,483,759 paired-end reads at 2 × 150 bp per sample after quality trimming and filtering. De novo transcriptomes were co-assembled using rnaSPAdes v3.13.1 (*80*), producing 520,555 representative transcripts from 496,603 putative genes; genes were estimated by rnaSPAdes.

We used TransDecoder v5.5.0 (*81*) for gene modeling in a multistep process to minimize false positives. In particular, we used the following procedure: (i) *TransDecoder.LongOrfs*, with transcript-to-gene mappings assigned by rnaSPAdes, to generate putative open reading frames (ORFs); (ii) *hmmscan* (hmmer v3.3.1 suite) (*82*) to identify protein domains using the PFAM v33.1 and TIGRFAM v15.0 databases; (iii) Diamond v0.9.30.131 (*83*) *blastp* against all *Scleractinia* (stony corals) proteomes available in NCBI (GCA_002571385.1, GCF_002042975.1, GCA_003704095.1, GCF_004143615.1, GCF_002571385.1, GCF_003704095.1, and GCF_000222465.1); and (iv) *TransDecoder.Predict* with the putative ORFs from (i), the protein domains from (ii), and the alignments from (iii) using the --single_best_only argument. This procedure generated a single ORF per transcript to yield 130,183 ORFs from 114,118 genes.

High-quality genes were annotated by using Diamond's *blastp* against NCBI's *nr* database (v2020.04.01), and taxonomic lineages were extrapolated from NCBItaxid using the *get_taxonomy_lineage_from_identifier* function from soothsayer v2020.08.24 (https://github.com/jolespin/soothsayer) with ete3 backend (*84*). PhyloDB v1.076 was used for additional annotations such as Kyoto Encyclopedia of Genes and Genomes ortholog assignments.

Orthogroups were identified using OrthoFinder v2.4.0 (*85*) with the high-quality proteins generated from our TransDecoder procedure and all of the *Scleractinia* proteomes listed previously. Annotations for orthogroups were assigned by using the most common organism-agnostic annotation within the grouping.

We assessed differential expression through a comparative genomics perspective for increased ecological interpretability. We aggregated the counts for each orthogroup to generate an orthogroup expression table with 27,140 orthogroup features. We filtered the expression table to include only orthogroups that were in at least 95% of the samples to yield a filtered count table with 17,755 orthogroup features.

For differential expression analysis, we used the 17,755 orthogroup set with the glmFIT and glmLRT models from edgeR v3.28.0 (*86*) and visualized the distributions and differentially expressed groups (DEGs) using *plot_volcano* from the *soothsayer* Python package (https://github.com/jolespin/soothsayer) (fig. S6). To build our design matrix for the generalized linear models, we use a global categorical approach where we used each (time point,

treatment, and temperature) grouping as a category and built a binary matrix. We estimated dispersion using the global dataset (not including T0 samples as they were not applicable) using estimateGLMCommonDisp and estimateGLMTagwiseDisp. This model structure allowed us to address variance between conditions while also providing a means to do time-specific contrasts.

Each condition had at least three biological replicates. The conditions investigated were the following using a threshold of FDR < 0.05: (i) placebo (26°C) versus placebo (30°C) (T2: 2294 DEGs; T3: 2795 DEGs); (ii) BMC (26°C) versus BMC (30°C) (T2: 35 DEGs; T3: 1426 DEGs); (iii) placebo (30°C) versus BMC (30°C) (T2: 0 DEGs; T3: 169 DEGs); and (iv) placebo (26°C) versus BMC (26°C) (T2: 2371 DEGs; T3: 0 DEGs) (table S2). As only minor differences were seen between BMC 30°C and placebo 30°C in T2, we focused on T3 transcriptome responses. For this, a graphic showing the DEGs with significant difference (FDR $P < 0.05$) between the condition BMC 30°C and placebo 30°C at T3 highlighted in the discussion was generated using the $\log_2$ FC and presented in Fig. 3, highlighting the mechanisms that might be involved in coral survivorship given by BMC treatment. The $\log_2$ FC and FDR values for each orthogroups and condition investigated (described in the paragraph above) can be found on table S2.

Metabolomic analysis

Fragments from each coral sample produced (300.00 mg) were homogenized with 80% methanol (1.50 ml) using zirconia bids and sonicated for 8 min at room temperature. The extraction mixtures were centrifuged at 10,000*g* for 10 min at 4°C, and the supernatants were concentrated to dryness under vacuum. This procedure was repeated three times for maximum recovery. The residues were resuspended in methanol-d_4 (200.00 ml) for NMR data acquisition, using 3-mm tubes and a 600-MHz Bruker Avance III equipped with a 5-mm TCI H-C/N-D cryoprobe and a SampleJet autosampler cooled samples to 6°C while waiting in the queue. The one-dimensional (1D) spectra (noesypr1D) experiment was used to assess the metabolomic profile of the dataset, and 2D experiments heteronuclear single-quantum coherence and heteronuclear multiple-bond correlation (hsqcedetgpsisp2.2 and hmbcetgpl3nd, respectively) were used to confirm the identity of key compounds. Quality control samples were included, and they have shown to be according to the expected. The spectra were processed using NMRPipe and imported into MATLAB for normalization, scaling, and multivariate analysis, using an in-house toolbox [developed in the Edison laboratory (*70*)]. The PLS-DA analysis was done using the 1D NMR spectra in full resolution, and the boxplot was constructed using the area under the curve of the peaks related to DMSP (2.88 ppm) and DMSO (2.58 ppm). The statistical difference of DMSP and DMSO between sampling times and treatments was assessed with an independent samples *t* test.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/7/33/eabg3088/DC1

REFERENCES AND NOTES

1. T. P. Hughes, K. D. Anderson, S. R. Connolly, S. F. Heron, J. T. Kerry, J. M. Lough, A. H. Baird, J. K. Baum, M. L. Berumen, T. C. Bridge, D. C. Claar, C. M. Eakin, J. P. Gilmour, N. A. J. Graham, H. Harrison, J. A. Hobbs, A. S. Hoey, M. Hoogenboom, R. J. Lowe, M. T. McCulloch, J. M. Pandolfi, M. Pratchett, V. Schoepf, G. Torda, S. K. Wilson, Spatial and temporal patterns of mass bleaching of corals in the Anthropocene. *Science* **359**, 80–83 (2018).
2. M. J. H. Van Oppen, R. D. Gates, L. L. Blackall, N. Cantin, L. J. Chakravarti, W. Y. Chan, C. Cormick, A. Crean, K. Damjanovic, H. Epstein, P. L. Harrison, T. A. Jones, M. Miller, R. J. Pears, L. M. Peplow, D. A. Raftos, B. Schaffelke, K. Stewart, G. Torda, D. Wachenfeld, A. R. Weeks, H. M. Putnam, Shifting paradigms in restoration of the world' s coral reefs. *Glob. Chang. Biol.* **23**, 1–12 (2017).
3. R. S. Peixoto, M. Sweet, D. G. Bourne, Customized medicine for corals. *Front. Mar. Sci.* **6**, 686 (2019).
4. T. C. LaJeunesse, J. E. Parkinson, P. W. Gabrielson, H. J. Jeong, J. D. Reimer, C. R. Voolstra, S. R. Santos, Systematic revision of symbiodiniaceae highlights the antiquity and diversity of coral endosymbionts. *Curr. Biol.* **28**, 2570–2580.e6 (2018).
5. L. Muscatine, J. W. Porter, Reef corals: Mutualistic symbioses adapted to nutrient-poor environments. *Bioscience* **27**, 454–460 (1977).
6. N. Rädecker, C. Pogoreutz, H. M. Gegner, A. Cárdenas, F. Roth, J. Bougoure, P. Guagliardo, C. Wild, M. Pernice, J. Raina, A. Meibom, C. R. Voolstra, Heat stress destabilizes symbiotic nutrient cycling in corals. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2022653118 (2021).
7. F. Rohwer, V. Seguritan, F. Azam, N. Knowlton, Diversity and distribution of coral-associated bacteria. *Mar. Ecol. Prog. Ser.* **243**, 1–10 (2002).
8. R. Littman, B. L. Willis, D. G. Bourne, Metagenomic analysis of the coral holobiont during a natural bleaching event on the Great Barrier Reef. *Environ. Microbiol. Rep.* **3**, 651–660 (2011).
9. C. Jaspers, S. Fraune, A. E. Arnold, D. J. Miller, T. C. G. Bosch, C. R. Voolstra, Resolving structure and function of metaorganisms through a holistic framework combining reductionist and integrative approaches. *Fortschr. Zool.* **133**, 81–87 (2019).
10. R. S. Peixoto, P. M. Rosado, D. C. A. Leite, A. S. Rosado, D. G. Bourne, Beneficial microorganisms for corals (BMC): Proposed mechanisms for coral health and resilience. *Front. Microbiol.* **8**, 341 (2017).
11. R. S. Peixoto, M. Sweet, H. D. M. Villela, P. Cardoso, T. Thomas, C. R. Voolstra, L. Høj, D. G. Bourne, Coral probiotics: Premise, promise, prospects. *Annu. Rev. Anim. Biosci.* **9**, 265–288 (2021).
12. M. Ziegler, F. O. Seneca, L. K. Yum, S. R. Palumbi, C. R. Voolstra, Bacterial community dynamics are linked to patterns of coral heat tolerance. *Nat. Commun.* **8**, 14213 (2017).
13. C. Bang, T. Dagan, P. Deines, N. Dubilier, W. J. Duschl, S. Fraune, U. Hentschel, H. Hirt, N. Hülter, T. Lachnit, D. Picazo, L. Pita, C. Pogoreutz, N. Rädecker, M. M. Saad, R. A. Schmitz, H. Schulenburg, C. R. Voolstra, N. Weiland-Bräuer, M. Ziegler, T. C. G. Bosch, Metaorganisms in extreme environments: Do microbes play a role in organismal adaptation? *Fortschr. Zool.* **127**, 1–19 (2018).
14. M. Ziegler, C. G. B. Grupstra, M. M. Barreto, M. Eaton, J. BaOmar, K. Zubier, A. Al-Sofyani, A. J. Turki, R. Ormond, C. R. Voolstra, Coral bacterial community structure responds to environmental change in a host-specific manner. *Nat. Commun.* **10**, 1–11 (2019).
15. C. R. Voolstra, M. Ziegler, Adapting with microbial help: Microbiome flexibility facilitates rapid responses to environmental change. *Bioessays* **42**, 2000004 (2020).
16. L. Reshef, O. Koren, Y. Loya, I. Zilber-Rosenberg, E. Rosenberg, The coral probiotic hypothesis. *Environ. Microbiol.* **8**, 2068–2073 (2006).
17. S. J. Robbins, C. M. Singleton, C. X. Chan, L. F. Messer, A. U. Geers, H. Ying, A. Baker, S. C. Bell, K. M. Morrow, M. A. Ragan, D. J. Miller, S. Forêt; ReFuGe2020 Consortium, C. R. Voolstra, G. W. Tyson, D. G. Bourne, A genomic view of the reef-building coral Porites lutea and its microbial symbionts. *Nat. Microbiol.* **4**, 2090–2100 (2019).
18. P. M. Rosado, D. C. A. Leite, G. A. S. Duarte, R. M. Chaloub, G. Jospin, U. N. Rocha, J. P. Saraiva, F. Dini-Andreote, J. A. Eisen, D. G. Bourne, R. S. Peixoto, Marine probiotics: Increasing coral resistance to bleaching through microbiome manipulation. *ISME J.* **13**, 921–936 (2019).
19. A. D. Williams, B. E. Brown, L. Putchim, M. J. Sweet, Age-related shifts in bacterial diversity in a reef coral. *PLOS ONE* **10**, e0144902 (2015).
20. N. S. Webster, T. B. H. Reusch, Microbial contributions to the persistence of coral reefs. *ISME J.* **11**, 2167–2174 (2017).
21. J. R. Zaneveld, R. Mcminds, R. V. Thurber, Stress and stability: Applying the Anna Karenina principle to animal microbiomes. *Nat. Microbiol.* **2**, 1–8 (2017).
22. D. P. Silva, G. Duarte, H. D. M. Villela, H. F. Santos, P. M. Rosado, J. G. Rosado, A. S. Rosado, E. M. Ferreira, A. U. Soriano, R. S. Peixoto, Adaptable mesocosm facility to study oil spill impacts on corals. *Ecol. Evol.* **9**, 5172–5185 (2019).
23. D. P. Silva, H. D. M. Villela, H. F. Santos, G. A. S. Duarte, J. R. Ribeiro, A. M. Ghizelini, C. L. S. Vilela, P. M. Rosado, C. S. Fazolato, E. P. Santoro, F. L. Carmo, D. S. Ximenes, A. U. Soriano, C. T. C. C. Rachid, R. L. V. Thurber, R. S. Peixoto, Multi-domain probiotic consortium as an alternative to chemical remediation of oil spills at coral reef and adjacent sites. *Microbiome* **9**, 118 (2021).
24. Y. Zhang, Q. Yang, J. Ling, L. Long, H. Huang, J. Yin, M. Wu, X. Tang, X. Lin, Y. Zhang, J. Dong, Shifting the microbiome of a coral holobiont and improving host physiology by inoculation with a potentially beneficial bacterial consortium. *BMC Microbiol.* **21**, 130 (2021).
25. C. R. Voolstra, C. B. López, G. Perna, A. Cárdenas, B. C. C. Hume, N. Rädecker, D. J. Barshis, Standardized short-term acute heat stress assays resolve historical differences in coral thermotolerance across microhabitat reef sites. *Glob. Change Biol* **26**, 4328–4343 (2020).

26. Y. Ben-Haim, F. L. Thompson, C. C. Thompson, M. C. Cnockaert, B. Hoste, J. Swings, E. Rosenberg, *Vibrio coralliilyticus* sp. nov., a temperature-dependent pathogen of the coral *Pocillopora damicornis*. *Int. J. Syst. Evol. Microbiol.* **53**, 309–315 (2003).
27. N. J. Alves, O. S. M. Neto, B. S. O. Silva, R. L. De Moura, R. B. Francini-Filho, C. B. Castro, R. Paranhos, B. C. Bitner-Mathé, R. H. Kruger, A. C. P. Vicente, C. C. Thompson, F. L. Thompson, Diversity and pathogenic potential of Vibrios isolated from Abrolhos Bank corals. *Environ. Microbiol. Rep.* **2**, 90–95 (2010).
28. A. P. Gasch, P. T. Spellman, C. M. Kao, O. C. Harel, M. B. Eisen, G. Storz, D. Botstein, P. O. Brown, Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* **11**, 4241–4257 (2000).
29. G. Dixon, E. Abbott, M. Matz, Meta-analysis of the coral environmental stress response: Acropora corals show opposing responses depending on stress intensity. *Mol. Ecol.* **29**, 2855–2870 (2020).
30. C. A. Morgans, J. Y. Hung, D. G. Bourne, K. M. Quigley, Symbiodiniaceae probiotics for use in bleaching recovery. *Restor. Ecol.* **28**, 282–288 (2020).
31. H. F. Santos, G. A. S. Duarte, C. T. C. C. Rachid, R. M. Chaloub, E. N. Calderon, L. F. B. Marangoni, A. Bianchini, A. H. Nudi, F. L. Carmo, J. D. van Elsas, A. S. Rosado, C. B. Castro, R. S. Peixoto, Impact of oil spills on coral reefs can be reduced by bioremediation using probiotic microbiota. *Sci. Rep.* **5**, 1–11 (2015).
32. W. Pootakham, W. Pootakham, W. Mhuantong, T. Yoocha, L. Putchim, N. Jomchai, C. Sonthirod, C. Naktang, W. Kongkachana, S. Tangphatsornruang, Heat-induced shift in coral microbiome reveals several members of the Rhodobacteraceae family as indicator species for thermal stress in *Porites lutea*. *Microbiol. Open* **12**, 935 (2019).
33. D. C. Leite, P. Leão, A. G. Garrido, U. Lins, H. F. Santos, D. O. Pires, C. B. Castro, J. D. van Elsas, C. Zilberberg, A. S. Rosado, R. S. Peixoto, Broadcast spawning coral *Mussismilia hispida* can vertically transfer its associated bacterial core. *Front. Microbiol.* **8**, 176 (2017).
34. D. C. Leite, J. F. Salles, E. N. Calderon, C. B. Castro, A. Bianchini, J. A. Marques, J. D. van Elsas, R. S. Peixoto, Coral bacterial-core abundance and network complexity as proxies for anthropogenic pollution. *Front. Microbiol.* **9**, 833 (2018).
35. D. C. Leite, J. F. Salles, E. N. Calderon, J. D. van Elsas, R. S. Peixoto, Specific plasmid patterns and high rates of bacterial co-occurrence within the coral holobiont. *Ecol. Evol.* **8**, 1818–1832 (2018).
36. H. Villela, Microbiome flexibility provides new perspectives in coral research. *Bioessays* **42**, e2000088 (2020).
37. M. Sweet, A. Burian, J. Fifer, M. Bulling, D. Elliott, L. Raymundo, Compositional homogeneity in the pathobiome of a new, slow-spreading coral disease. *Microbiome* **7**, 139 (2019).
38. H. F. Santos, F. L. Carmo, J. E. S. Paes, A. S. Rosado, R. S. Peixoto, Bioremediation of mangroves impacted by petroleum. *Water Air Soil Pollut.* **216**, 329–350 (2011).
39. P. Pandey, S. Bisht, A. Sood, A. Aeron, G. D. Sharma, D. K. Maheshwari, Consortium of plant growth-promoting-bacteria: Future perspectives in agriculture, in *Bacteria in Agrobiology: Plant Probiotics*, D. K. Maheshwari, Ed. (Springer, 2012), pp 185–200.
40. R. S. C. de Souza, J. S. L. Armanhi, P. Arruda, From microbiome to traits: Designing synthetic microbial communities for improved crop resiliency. *Front. Plant Sci.* **11**, 1179 (2020).
41. M. J. Sweet, B. E. Brown, R. P. Dunne, I. Singleton, M. Bulling, Evidence for rapid, tide-related shifts in the microbiome of the coral *Coelastrea aspera*. *Coral Reefs* **36**, 815–828 (2017).
42. S. Takahashi, H. Bauwe, M. R. Badger, Impairment of the photorespiratory pathway accelerates photoinhibition of photosystem II by suppression of repair but not acceleration of damage processes in Arabidopsis. *Plant Physiol.* **144**, 487–494 (2007).
43. M. P. Lesser, W. R. Stochaj, D. W. Tapley, J. M. Shick, Bleaching in coral reef anthozoans: Effects of irradiance, ultraviolet radiation, and temperature on the activities of protective enzymes against active oxygen. *Coral Reefs* **8**, 225–232 (1990).
44. T. Bieri, M. Onishi, T. Xiang, A. R. Grossman, J. R. Pringle, Relative contributions of various cellular mechanisms to loss of algae during cnidarian bleaching. *PLOS ONE* **11**, e0152693 (2016).
45. S. R. Dunn, M. Pernice, K. Green, O. Hoegh-guldberg, S. G. Dove, Thermal stress promotes host mitochondrial degradation in symbiotic cnidarians: Are the batteries of the reef going to run out? *PLOS ONE* **7**, e39024 (2012).
46. T. G. Cross, D. S. Toellner, N. V. Henriquez, E. Deacon, M. Salmon, J. M. Lord, Serine/threonine protein kinases and apoptosis. *Exp. Cell Res.* **256**, 34–41 (2000).
47. M. Kurokawa, S. Kornbluth, Caspases and kinases in a death grip. *Cell* **138**, 838–854 (2009).
48. R. Yehuda, C. W. Hoge, A. C. McFarlane, E. Vermetten, R. A. Lanius, C. M. Nievergelt, S. E. Hobfoll, K. C. Koenen, T. C. Neylan, S. E. Hyman, Post-traumatic stress disorder. *Nat. Rev. Dis. Primers.* **1**, 15057 (2015).
49. V. M. Weiss, Cellular mechanisms of Cnidarian bleaching: Stress causes the collapse of symbiosis. *J. Experim. Biol.* **211**, 3059–3066 (2008).
50. L. Thomas, S. R. Palumbi, The genomics of recovery from coral bleaching. *Proc. R. Soc. B* **284**, 20171790 (2017).
51. J. H. Pinzón, B. Kamel, C. A. Burge, C. D. Harvell, M. Medina, E. Weil, L. D. Mydlarz, Whole transcriptome analysis reveals changes in expression of immune-related genes during and after bleaching in a reef-building coral. *R. Soc. Open Sci.* **21**, 40214 (2015).
52. R. Savary, D. J. Barshis, C. R. Voolstra, A. Cárdenas, N. R. Evensen, G. B. Prandi, M. Fine, A. Meibom, Fast and pervasive transcriptomic resilience and acclimation of extremely heat-tolerant coral holobionts from the northern Red Sea. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2023298118 (2021).
53. M. Garren, K. Son, J.-B. Raina, R. Rusconi, F. Menolascina, O. H. Shapiro, J. Tout, D. G. Bourne, J. R Seymour, R. Stocker, A bacterial pathogen uses dimethylsulfoniopropionate as a cue to target heat-stressed corals. *ISME J.* **8**, 999–1007 (2014).
54. J. B. Raina, D. Tapiola, C. A. Motti, S. Foret, T. Seemann, J. Tebben, B. L. Willis, D. G. Bourne, Isolation of an antimicrobial compound produced by bacteria associated with reef-building corals. *PeerJ* **4**, e2275 (2016).
55. J. B. Raina, P. L. Clode, S. Cheong, J. Bougoure, M. R. Kilburn, A. Reeder, S. Forêt, M. Stat, V. Beltran, P. T. Hall, D. Tapiolas, C. M. Motti, B. Gong, M. Pernice, C. E. Marjo, J. R. Seymour, B. L. Willis, D. G. Bourne, Subcellular tracking reveals the location of dimethylsulfoniopropionate in microalgae and visualises its uptake by marine bacteria. *eLife* **6**, e23008 (2017).
56. J. S. Wirth, T. Wang, Q. Huang, R. H. White, W. B. Whitman, Dimethylsulfoniopropionate sulfur and methyl carbon assimilation in *Ruegeria* species. *MBio* **11**, 2 (2020).
57. N. Miura, K. Motone, T. Takagi, S. Aburaya, S. Watanabe, W. Aoki, M. Ueda, *Ruegeria sp.* strains isolated from the reef-building coral *Galaxea fascicularis* inhibit growth of the temperature-dependent pathogen Vibrio coralliilyticus. *J. Mar. Biotechnol.* **21**, 1–8 (2019).
58. D. K. Dahiya, R. Malik, M. Puniya, U. K. Shandilya, T. Dhewa, N. Kumar, S. Kumar, A. K. Puniya, P. Shukla, Gut microbiota modulation and its relationship with obesity using prebiotic fibers and probiotics: A review. *Front. Microbiol.* **8**, 563 (2017).
59. D. Shin, S. Y. Chang, P. Bogere, K. H. Won, J. Y. Choi, Y. J. Choi, H. K. Lee, J. Hur, B. Y. Park, Y. Kim, J. Heo, Beneficial roles of probiotics on the modulation of gut microbiota and immune response in pigs. *PLOS ONE* **14**, e0220843 (2019).
60. C. M. Eakin, J. A. Morgan, S. F. Heron, T. B. Smith, G. Liu, L. A. Filip, B. Baca, E. Bartels, C. Bastidas, C. Bouchon, M. Brandt, A. W. Bruckner, L. B. Williams, A. Cameron, B. D. Causey, M. Chiappone, T. R. L. Christensen, M. J. C. Crabbe, O. Day, E. de la Guardia, G. D. Pulido, D. DiResta, D. L. G. Agudelo, D. S. Gilliam, R. N. Ginsburg, S. Gore, H. M. Guzmán, J. C. Hendee, E. A. H. Delgado, E. Husain, C. F. G. Jeffrey, R. J. Jones, E. J. Dahlgren, L. S. Kaufman, D. I. Kline, P. A. Kramer, J. C. Lang, D. Lirman, J. Mallela, C. Manfrino, J. P. Maréchal, K. Marks, J. Mihaly, W. J. Miller, E. M. Mueller, E. M. Muller, C. A. O. Toro, H. A. Oxenford, D. P. Taylor, N. Quinn, K. B. Ritchie, S. Rodríguez, A. R. Ramírez, S. Romano, J. F. Samhouri, J. A. Sánchez, G. P. Schmahl, B. V. Shank, W. J. Skirving, S. C. C. Steiner, E. Villamizar, S. M. Walsh, C. Walter, E. Weil, E. H. Williams, K. W. Roberson, Y. Yusuf, Caribbean corals in crisis: Record thermal stress, bleaching, and mortality in 2005. *PLOS ONE* **5**, e13969 (2010).
61. K. D. Bahr, K. S. Rodgers, P. L. Jokiel, Impact of three bleaching events on the reef resiliency of Kāne'ohe Bay, Hawai'i. *Front. Microbiol.* **4**, 398 (2017).
62. G. A. S. Duarte, H. D. M. Villela, M. Deocleciano, D. Silva, A. Barno, P. M. Cardoso, C. L. S. Vilela, P. Rosado, C. S. M. A. Messias, M. A. Chacon, E. P. Santoro, D. B. Olmedo, M. Szpilman, L. A. Rocha, M. Sweet, R. S. Peixoto, Heat waves are a major threat to turbid coral reefs in Brazil. *Front. Mar. Sci.* **7**, 179 (2020).
63. National Academies of Science, Engineering and Medicine, *A Decision Framework for Interventions to Increase the Persistence and Resilience of Coral Reefs* (National Academies Press, 2019).
64. J. Kleypas, K. Son, J.-B. Raina, R. Rusconi, F. Menolascina, O. H. Shapiro, J. Tout, D. G. Bourne, J. R. Seymour, R. Stocker, Designing a blueprint for coral reef survival. *Biol. Conserv.* **257**, 109107 (2021).
65. M. Giambiagi-Marval, M. A. Mafra, E. G. C. Penido, M. C. F. Bastos, Distinct groups of plasmids correlated with bacteriocin production in *Staphylococcus aureus*. *J. Gen. Microbiol.* **136**, 1591–1599 (1990).
66. J. Vidal-Dupiol, O. Ladrière, A. L. Meistertzheim, L. Fouré, M. Adjeroud, G. Mitta, Physiological responses of the scleractinian coral *Pocillopora damicornis* to bacterial stress from *Vibrio coralliilyticus*. *J. Experim. Biol.* **214**, 1533–1545 (2011).
67. J. Vidal-Dupiol, O. Ladrière, D. D. Garzón, P. E. Sautière, A. L. Meistertzheim, E. Tambutté, S. Tambutté, D. Duval, L. Fouré, M. Adjeroud, G. Mitta, Innate immune responses of a scleractinian coral to vibriosis. *J. Biol. Chem.* **286**, 22688–22698 (2011).
68. S. E. Cole, F. J. Lariviere, C. N. Merrikh, M. J. Moore, A convergence of rRNA and mRNA quality control pathways revealed by mechanistic analysis of nonfunctional rRNA decay. *Mol. Cell* **34**, 40–50 (2009).
69. T. Hall, BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* **41**, 95–98 (1999).
70. S. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
71. A. F. Andersson, M. Lindberg, H. Jakobsson, P. Nyrén, L. Engstrand, Comparative analysis of human gut microbiota by barcoded pyrosequencing. *PLOS ONE* **3**, 2836 (2008).

72. N. A. Bokulich, B. D. Kaehler, J. R. Rideout, M. Dillon, E. Bolyen, R. Knight, G. A. Huttley, J. G. Caporaso, Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome* **6**, 90 (2018).
73. G. Henderson, P. Yilmaz, S. Kumar, R. J. Forster, W. J. Kelly, S. C. Leahy, L. L. Guan, P. H. Janssen, Improved taxonomic assignment of rumen bacterial 16S rRNA sequences using a revised SILVA taxonomic framework. *PeerJ* **7**, e6496 (2019).
74. K. Katoh, K. I. H. Kuma, T. Miyata, Mafft version 5: Improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511–518 (2005).
75. P. J. Mcmurdie, S. Holmes, Phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLOS ONE* **8**, 61217 (2013).
76. H. Wickham, R. Francois, L. Henry, K. Müller, Dplyr: A Grammar of Data Manipulation, R Package Version 0.7.3 (2017).
77. H. Wickham, ggplot2. *Wiley Interdiscip. Rev. Comput. Stat.* **3**, 180–185 (2011).
78. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
79. J. Oksanen, F. G. Blanchet, M. Friendly, R. Kindt, P. Legendre, D. McGlinn, P. R. Minchin, R. B. O'Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens, E. Szoecs, H. Wagne, Vegan: Community ecology package. R package version 2.0-10 (2019).
80. E. Bushmanova, D. Antipov, A. Lapidus, A. D. Prjibelski, rnaSPAdes: A de novo transcriptome assembler and its application to RNA-Seq data. *Gigascience* **8**, giz100 (2019).
81. B. J. Haas, A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, M. B. Couger, D. Eccles, B. Li, M. Lieber, M. D. M. Manes, M. Ott, J. Orvis, N. Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C. N. Dewey, R. Henschel, R. D. Le Duc, N. Friedman, A. Regev, De novo transcript sequence reconstruction from RNA-Seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
82. S. R. Eddy, Accelerated profile HMM searches. *PLOS Comput. Biol.* **7**, e1002195 (2011).
83. B. Buchfink, C. Xie, D. H. Huson, Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
84. J. Huerta-Cepas, F. Serra, P. Bork, ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
85. D. M. Emms, S. Kelly, OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
86. M. D. Robinson, D. J. Mccarthy, G. K. Smyth, edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

Acknowledgments: We thank D. Pimenta and R. Grilo for logistical and sampling support in the Maraú peninsula and A. O'rourke for technical support and discussions. We also thank the CCRC NMR Facility, especially A. S. Edison, for instrumentation for NMR data collection.
Funding: This research project won the Great Barrier Reef Foundation's Out of the Blue Box Reef Innovation Challenge People's Choice Award supported by The Tiffany & Co. Foundation. C.R.V. was supported through the Deutsche Forschungsgemeinschaft (DFG; German Research Foundation) Project Numbers 433042944 and 458901010. R.S.P. was supported through KAUST grant number BAS/1/1095-01-01 and the Rio de Janeiro Marine Aquarium Research Center. E.P.S. received support from the Graduate Programs of Science (Microbiology) and Plant Biotechnology and Bioprocess Engineering (PBV)/Federal University of Rio de Janeiro, the Brazilian Government Research Agency CAPES, and the CAPES PRINT international mobility grant.

Author contributions: E.P.S., H.D.M.V., G.A.S.D., and R.S.P conceived and designed the study; G.A.S.D. and E.P.S. collected the coral samples; E.P.S., C.S.M.A.M., H.D.M.V., C.L.S.V., J.G.R., P.M.C., P.M.R., J.M.A., and G.A.S.D. performed the BMC selection and mesocosm experiment; R.M.B. performed the metabolomic analysis; J.L.E., C.L.D., K.E.N., L.M.P., C.R.V., and G.P. were involved with DNA- or RNA-seq, bioinformatics, and statistical analyses; E.P.S., J.L.E., C.L.D., C.R.V., L.M.P., M.J.S., A.S.R., A.M., R.M.B., and R.S.P. analyzed and interpreted the data; E.P.S., C.R.V., and R.S.P. wrote the manuscript; J.L.E., C.L.D., M.J.S., H.D.M.V., A.S.R., A.M., and R.M.B. were involved in the critical revision of the manuscript; and R.S.P., K.E.N., and C.R.V. provided financial support. All author(s) read and approved the final manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. *M. hispida* transcripts, gene models, and annotations are available at https://doi.org/10.6084/m9.figshare.13344758.v1.

Submitted 23 December 2020
Accepted 24 June 2021
Published 13 August 2021
10.1126/sciadv.abg3088

Citation: E. P. Santoro, R. M. Borges, J. L. Espinoza, M. Freire, C. S. M. A. Messias, H. D. M. Villela, L. M. Pereira, C. L. S. Vilela, J. G. Rosado, P. M. Cardoso, P. M. Rosado, J. M. Assis, G. A. S. Duarte, G. Perna, A. S. Rosado, A. Macrae, C. L. Dupont, K. E. Nelson, M. J. Sweet, C. R. Voolstra, R. S. Peixoto, Coral microbiome manipulation elicits metabolic and genetic restructuring to mitigate heat stress and evade mortality. *Sci. Adv.* **7**, eabg3088 (2021).

4 - CRITICAL OVERVIEW

4.1 - Development of open-source software and algorithms

The primary objective of the research for this dissertation was to develop open-source human interpretable artificial intelligence methods and use these tools to characterize various types of microbial-related diseases from the perspective of systems biology. The secondary objective was to develop these methodologies with the intention of domain-agnostic generalizability; that is, although these algorithms were developed for biotechnology they can be applied towards other sciences seamlessly without modification. The approach taken when designing these algorithms and frameworks was to address a broad set of analytical questions and for synergy between methodologies.

The methods developed are implemented in the Python programming language and open-sourced through my *Soothsayer Ecosystem* (https://github.com/jolespin) which includes the following packages: 1) *Soothsayer*; 2) *Ensemble NetworkX*; 3) *Hive NetworkX*; 4) *Compositional*; and 5) *GenoPype*. *Soothsayer* is high-level data science package, used in **Publications I-V**, that provides analytical methods including the *Clairvoyance* feature selection algorithm and the hierarchical ensemble of classifiers model framework described in **Publication II**. The *Ensemble NetworkX* package builds on the existing *NetworkX* package (Hagberg et al., 2008) while adding implementations for ensemble networks, differential ensemble networks, sample-specific perturbation networks, and categorical feature engineering as described in **Publication III** and **IV**. The *Hive NetworkX* is another extension of *NetworkX* but specializes in the visualization of highly complex networks using hive plots (Krzywinski et al., 2012) described in

Publications I, III, and **IV**. The *Compositional* package implements compositional data transformations and proportionality metrics for network analysis which are used as a backend in *EnsembleNetworkX* described in **Publication I, III,** and **IV**. Lastly, the *GenoPype* package which was developed to efficiently build complex computational pipelines with a standardized file structure, checkpoints, logs, and emphasizing reproducibility.

Table 4 – Open-sourced software packages developed for dissertation

Package	Description	URL
Soothsayer	High-level (bio-)informatics package	github.com/jolespin/soothsayer
Ensemble NetworkX	Methods for ensemble network analysis	github.com/jolespin/ensemble_networkx
Hive NetworkX	Data structures and visualizations for hive plots	github.com/jolespin/hive_networkx
Compositional	Methods for compositional data analysis	github.com/jolespin/compositional
GenoPype	Architecture for building computational pipelines	github.com/jolespin/genopype

4.2 – Applications of weighted association networks applied to compositional data in biology

The aim for **Publication I** was to serve as a prelude for the research articles that comprise this dissertation and set a foundation for the caveats my methodologies are designed to overcome. In particular, the first objective of this review was to address how NGS data is compositional (and not optional), what types of errors arise when not factoring in compositionality, and why analyzing NGS from the perspective of CoDA is paramount for biological interpretation. The second objective of this review was to introduce the dimensionality limitations that are inherent in biological datasets, how these problems are

exacerbated when investigating problems from a network perspective, and how to factor in compositionality when conducting experiments *in silico* at the systems level. The foundations set for CoDA, data dimensionality, and network analysis are critical for the XAI methodologies that were developed in this dissertation for investigating microbial-related diseases such as antibiotic resistance, acute malnutrition, and dental caries.

4.2.1 – Establishing a foundation for network analysis using next generation sequencing technologies

This review illustrated the utility of network theory in biological systems and investigated modern techniques while introducing researchers to frameworks for implementation. **Publication I** overviewed (1) CoDA principles, (2) compositionally-valid data transformations, and (3) network theory along with insight on a battery of network types including static-, temporal-, sample-specific-, and differential-networks (**Publication II** - Fig. 1,2). Further, the intention was not to provide a comprehensive overview of network methods, rather to introduce microbiology researchers to (semi)-unsupervised data-driven approaches for inferring latent structures that may give insight into biological phenomena or abstract mechanics of complex systems.

As this publication was a review, the methodology, results, and conclusions function synergistically as a meta-analysis. The basis of this meta-analysis was that systems researchers are not only investigating the abundance/depletion of features in relation to a specific condition but also the (inferred) interactions between features and implementation for large datasets must be pursued with diligence adhering to established principles. One way for such an investigation into these inferred interactions is by applying

network theory as the versatility of graphical abstractions using nodes, edges, and topological structure can be contextually applied to a wide array of problems (**Publication I** - Fig. 3). For instance, applications of network theory have been successful in several fields including studying plankton networks driving carbon export (Guidi et al., 2016), gene interactions related to weight physiology (Fuller et al., 2007), ecological shifts (Gomez et al., 2017) and metabolic potential (Espinoza et al., 2018) associated with carious lesions in children, and regulatory metabolic interactions in marine diatoms (Levering et al., 2017), and bacterial soil communities (Mandakovic et al., 2018). Many biological networks are composed of molecules such as DNA, RNA, proteins, and metabolites as the nodes, and edges between these nodes represent either curated or inferred interactions between them. Furthermore, advanced multi-omics approaches incorporating associations across modalities such as clinical tests, proteomics, amplicon, transcriptomics, cytokines, metabolomics, and lipidomics have begun to pave the way towards precision health using systems biology (Schüssler-Fiorenza Rose et al., 2019; Shomorony et al., 2020; Zhou et al., 2019). There are several approaches for network analysis in systems biology that each have their advantages and caveats. In conclusion, this complication of literature briefly describes the landscape of network methods but primarily guides the reader through the process of implementing association networks from NGS-derived datasets which are inherently compositional.

4.3 - Predicting antimicrobial mechanism-of-action from transcriptomes: A generalizable explainable artificial intelligence approach

The original aim of this project was to develop an efficient transcriptomic approach to dereplicate antibacterial extracts from uncultured soil bacteria, revealing which are chemically similar to known antibiotics. This approach quickly and inexpensively provided

the potential to find novel antimicrobials from new sources without wasting time and resources on already known antibacterial compounds; that is, obviating the bottleneck in identifying candidate MOA of novel compounds followed by prediction of whether these compounds hit a novel target or represent an entirely new MOA. The overarching goal was to resolve the dereplication bottleneck in natural product antibacterial drug discovery through a combination of *in silico* methodologies and empirical validation experiments.

Approximately 1,500 microorganism-produced compounds with demonstrated antimicrobial activity are commercially available. Despite this resource, the MOA and targets are unknown for most of these compounds. Therefore, the first aim of this research was to identify MOA and targets of known antimicrobials by transcriptome analysis. In doing so, we built a database of transcription profiles (i.e., gene expression) produced by antibiotics with known MOAs and targets and this served the training data resource for analysis of unknown compounds. This work profiled 41 clinically approved antibiotics and, in a collaborative effort, screened additional compounds obtained from the NIH National Center for Advancing Translational Sciences for antibacterial activity; all compounds that were modeled exhibited antibacterial activity (**Publication II** – Table 1). To produce transcriptome profiles, cultures of an *Escherichia coli* K12 mutant strain W0153 were challenged by the antimicrobial compound which were then assayed using NGS. Using the database of transcriptome profiles for antibiotics and their known MOA as training data, statistical methods were used to characterize each compound and XAI algorithms were developed to predict the MOA from unobserved compounds with high accuracy while also flagging compounds with potentially novel targets. The classification models and their ability to predict MOA and targets for these additional antimicrobial

compounds were validated using the most stringent of metrics (described below in 4.3.1). The profiles from these additional compounds, their MOAs, and targets were used to expand the size and utility of our antibiotic challenged transcriptomic database.

The second aim of this research was to identify MOA and targets of antimicrobials in extracts from uncultured bacteria. Uncultured bacteria represent an untapped source of chemical diversity, and these natural systems were leveraged for discovery of novel antibiotics. Similar to the pure compounds, these crude extracts were added to a growing culture of *E. coli* and RNA was collected for transcriptional profiling. The XAI models developed were calibrated to handle crude extracts and were used to predict the MOA of crude extracts with known antibiotic activity but unknown MOA. With this methodology, transcriptome profiles from extracts that clearly indicated known MOA can be included in the training data but, erring on the side of caution, were only included if empirically validated. Transcriptomes that were difficult to interpret, such as those indicating a novel MOA not represented by antibiotics in the training data can be further investigated as was the case with darobactin as this was confirmed in our sister studies (Imai et al., 2019; O'Rourke et al., 2020).

4.3.1 – Evaluating MOA prediction performance on unobserved compounds

Spearheading these aims was non-trivial. The most prominent obstacles encountered was how to properly evaluate the predictive performance on unobserved compounds and how to optimize this performance in the face of the aforementioned "curse of dimensionality". Many studies will use *K*-fold cross-validation, where the dataset is

(pseudo-)randomly split into *K* folds of training and testing data (e.g., 70% training/30% testing and repeat *K* times), to evaluate performance of a model on unobserved data and this typically scored using the average accuracy of each split. The problem with this approach in the **Publication II** dataset is the fact that the transcriptomic data was hierarchical with replicates. More specifically, there are multiple transcriptomes that represent a particular compound in the form of replicates. Therefore, if *K*-fold cross-validation was implemented then there would be an extremely high probability of having some representative of a compound in both the training and testing splits (in a dataset of this dimensionality). In this event, the model would artificially inflate the performance metrics and would not generalize to novel compounds since it has been trained on the same compounds it is testing. This hierarchy required one to implement a robust cross-validation and performance metric that could simulate the prediction of compounds that have not been observed previously by the model. An approach that abides by this stringency and is Leave Compound Out Cross-Validation (LCOCV) where all instances of a compound are reserved for a testing set and the remaining compounds are used for model training; thus, demonstrating predictive performance on unobserved compounds.

This modeling framework was also evaluated using external datasets. The first study evaluated was Hutter et al. which generated a database of *Bacillus subtilis* transcriptional responses to treatments of 37 well-characterized antibacterial compounds from different MOA which were used to build a support vector machine model to predict MOA of antibacterial compounds (Hutter et al., 2004). The training data from Hutter et al. 2004 was not published in any public database but the methodologies used were

reimplemented in **Publication II**. The second study evaluated was Zoffmann et al. 2019 which used a combination of transcriptomics and cell imaging data to predict 7 MOA in a different *E. coli* strain (BW25113) (Zoffmann et al., 2019). Zoffmann et al. did not publish the cell imaging data used to construct predictive models but did publish the counts from NCBI Gene Expression Omnibus consisting of E. coli BW25113 challenged with the 16 compounds; this data was modeled using CoHEC models. The third study evaluated was Zampieri et al. 2018 which used an iterative hypergeometric test to model metabolite responses of *Mycobacterium smegmatis* exposed to 62 compounds representing 18 MOAs. In all cases, the CoHEC models outperformed the models in the original studies and the results are detailed in **Publication II** – Table 3.

4.2.2 – *The dimensionality of transcriptomes with nested and imbalanced classes*

Regarding the dimensionality obstacles, not only did the number of features vastly exceed the number of observations but the number of observations per class were largely imbalanced, in this case, genes as features and transcriptomes as observations. The training dataset of **Publication II** consisted of 41 antibiotic compounds representing 6 MOA including inhibitors for protein synthesis (N_{Pure} = 9; N_{Crude} = 2), DNA synthesis (N_{Pure} = 10; N_{Crude} = 2), RNA synthesis (N_{Pure} = 4; N_{Crude} = 2), cell-wall synthesis (N_{Pure} = 12; N_{Crude} = 4), cell-membrane synthesis (N_{Pure} = 2; N_{Crude} = 0), and fatty-acid-synthesis synthesis (N_{Pure} = 3; N_{Crude} = 0) (**Publication II** – Table 1). These 41 compounds include an overlap of pure compounds and crude extracts that were processed using 3 biological replicates, though, some samples failed during library preparation or sequencing and were resequenced giving some compounds more than 3 replicates. With typical DGE

methodologies such as those implemented in *edgeR* (Robinson et al., 2010) or *DESeq2* (Love et al., 2014), the replicates would be fit to a model and the replicates would be reduced to a single log fold-change (logFC) and false discovery rate (FDR) per gene per replicate for a training data dimensionality of 41 compounds by ~4000 genes. While this is desirable during testing for enrichment or depletion of specific features relative to a hypothesis determined *a priori*, this is not advantageous in a machine learning setting where increased numbers of samples relative to features dramatically decreases the opportunity for model overfitting.

4.3.3 – *Enriching the gene feature set to remove genes not relevant to MOA prediction*

As the aim of this dissertation was to develop XAI methodologies that can be interpreted using domain knowledge, the first step towards addressing this dimensionality obstacle was by removing all genes that did not encode for proteins leaving 3065 protein-coding genes. The rationale for this was to exclude poorly characterized genes that may have created bottleneck in biological interpretation of the XAI models that could only be alleviated by empirical analysis of *E. coli* genes; a topic largely out of scope for this project. However, 3065 features is still vastly larger than the number of observations that can be acquired from this dataset (even after pairwise combinatorics as mentioned below in 4.3.4). The solution to overcome this dimensionality obstacle was to develop the *Clairvoyance* feature selection algorithm as a means for curating a gene set that could robustly discriminate the primary MOA of DGE profiles. The objective function implemented in *Clairvoyance* maximizes the accuracy of custom (or stochastic) cross-

validation pairs by iteratively enriching the subset of predictive features (e.g., genes) and, thereby, denoising the dataset with respect to a specific classification task and resulting in a smaller feature set with reduced potential for model overfitting. In the case of this study, the *Clairvoyance* algorithm iteratively refines the gene set for each sub-model to maximize the model's MOA classification accuracy of unobserved compounds which was provided as the test set of the LCOCV pairs.

4.3.4 – *Maximizing the number of observations while preserving information content*

The second action item to disarm this dimensionality complex was to address the observation to feature ratio as we only have 41 unique compounds and 3065 genes. As mentioned, the transcriptome samples are hierarchical with the lowest level being compounds and the next level being individual replicate transcriptomes. Therefore, this problem can be approached in several ways: 1) model the differential abundance of genes for grouped replicates challenged with a compound relative to grouped solvent negative controls (N=41 compounds); 2) consider each replicate individually relative to grouped solvent controls (N=235 transcriptomes); or 3) consider each pairwise combination of treatment compound and solvent replicates (N=713 pairwise DGE profiles) (**Publication II** – Table 1). The latter takes advantage of the natural hierarchy within the data by utilizing all the information contained within the samples and between the samples while also preserving variance that is critical for building generalizable models.

4.3.5 – *Using a hierarchical ensemble of classifiers to overcome class imbalance and predict MOA for unobserved compounds with high accuracy*

Lastly, the class imbalance needed to be addressed but this could not be accomplished using combinatorics as was done for the sample dimensionality nor feature selection as was used for the reduced gene set. The training data dimensionality was not ideal for even simple binary classification models, let alone 6 imbalanced classes, thus, it was not surprising to find that most traditional classification models performed poorly (<90% LCOCV accuracy) (**Publication II** – Table 2). The models evaluated included commonly used algorithms such as logistic regression (the highest performing of these traditional algorithms), random forest, K-nearest neighbors, support vector machines, naive bayes, AdaBoost, and even modern algorithms such as neural networks. Therefore, a model structure had to be devised that could handle this class imbalance on a case-by-case basis. With inspiration from the mechanisms of human cognition, this class imbalance was addressed by decomposing the complex task of multiclassification into a multilayered path of simple binary tasks which was achieved through the advent of a Hierarchical Ensemble of Classifiers (HEC) and the extend *Clairvoyance*-optimized HEC (CoHEC) models.

The basic HEC model approach implements a hierarchical ensemble of binary classifiers through a single graphical model with 3 degrees of flexibility for each sub-model decision node: (1) a custom feature set optimized for a simple binary classification task; (2) a unique classification algorithm with hyperparameters that most effectively discriminates the sub-model-specific decision paths; and (3) the relationship between sub-models can be data-driven or assigned *a priori*. The graphical structure of the MOA predictive CoHEC

model implemented in this study was entirely data-driven to demonstrate the autonomous abilities of the XAI methodology by solely using emergent patterns within the training dataset in relation to the labeled classes (**Publication II** – Fig. 1,2). In other words, the graphical structure or gene sets are not define *a priori* using curated databases or domain knowledge (although, this functionality is supported) and instead allow the data to guide such parameter choices.

Optimization of the gene feature set for each sub-model using *Clairvoyance* ($GeneSet_{yk}$ where k ranges from sub-models 1-5) boosted LCOCV accuracy substantially; between 10-23% in most cases and all cases resulting in left-out compound accuracies greater than 99%. Several estimators were evaluated, optimized, and tuned for each sub-classification task but logistic regression models were the exemplar in all cases. While a few genes are shared between various pairs of sub-models, none of the 399 unique genes from $GeneSet_{y1\ y5}$ used in the CoHEC model were universal to all sub-models reinforcing the notion that each sub-model is task specific (**Publication II** – Table 3). To showcase the advantage of combining feature selection with HEC models (i.e., CoHEC), *Clairvoyance* was also run using a traditional multiclass logistic regression to produce $GeneSet_{Multiclass}$. Of these 399 genes in our CoHEC model, there were 87 of the 98 genes (88.8%) in $GeneSet_{Multiclass}$ overlapping and, thus, demonstrating the ability of *Clairvoyance* to identify emergent patterns within the data despite different model architectures (**Publication II** – Table 3). However, the CoHEC model can exceed the performance of a multiclass model using the same base algorithm (e.g., logistic regression in this study) with only a fraction of the training data when using the same

input features. Interestingly, none of the MOA enzymatic targets were selected by *Clairvoyance* as discriminative features further endorsing our data-driven approach because the discriminating patterns were unknown *a priori*.

4.3.6 – *Identifying compounds with novel targets not previously characterized*

The final objective was for the methodology to be able to detect compounds with novel activity by using only their transcriptomic response. In collaboration with our extended team, we identified and characterized a novel antibiotic called darobactin which exhibits a novel MOA by targeting the BAM complex (Imai et al., 2019). The XAI methodology developed was able to flag novelty for transcriptomes of *E. coli* that have been challenged with darobactin. This was achieved by predicting the MOA for the transcriptomes using our CoHEC model but to then examine the variance of the resulting predictions (**Publication II** – Fig. 3). Upon implementing this variance comparison, it was possible to confidently suggest that darobactin exhibits an activity not observed by any of the original 6 MOA that were characterized in the study.

With the data transformation, feature selection, and hierarchical modeling frameworks developed, we were able to accomplish all our aims and objectives. Not only were methods developed that could alleviate this bottleneck in antibiotic discovery research, but the algorithms were designed in such a way that they could be generalized to a wide array of research topics as demonstrated in the Publication III childhood undernutrition study.

4.3.7 – *Interpreting CoHEC models in the context of antibiotic discovery*

Interpretability of trained models is paramount in XAI and CoHEC models provide great insight into the decision-making process. For instance, CoHEC models use a unique feature set for each of 5 sub-models (*y1*-*y5*) and these feature sets have very minimal overlap showcasing that each sub-model is engineered for a specific task (**Publication II** – Fig. 1C). Furthermore, CoHEC models produce an array of probabilities for each of the 5 sub-models with methods designed to calculate the probability for traversing each of the 10 decision paths and to visualize the predictions via decision graphs (**Publication II** – Fig. 2). In this case, the probabilities represent binary decision paths from each of the 5 logistic regression sub-models, though other algorithms for sub-models are supported, and the standard error is calculated for profiles grouped by LCOCV test set; that is, all associated pairwise DGE profiles corresponding to a compound in a LCOCV test set. These 10 probabilities computed by the CoHEC model on LCOCV test sets are machine informative as unsupervised analysis of these probabilities clusters compounds by MOA with statistically greater homogeneity than the input data of pairwise DGE profiles (**Publication II** – Fig. 3). The ability of CoHEC models to compute probabilities that can confidently cluster a compound with its respective producer-strain extract in a completely unsupervised setting provides a powerful avenue to dereplicate known compounds in high throughput.

Interpreting models based on gene expression data is difficult as this approach often captures downstream effects. Regardless, the decision graphs and sub-model gene coefficients are biologically relevant when evaluated via *Gene Set Enrichment Analysis*

(GSEA) (Subramanian et al., 2005). For instance, coefficient-ranked genes from sub-model *y2* (DNA-synthesis vs. *y4*) are enriched in both DNA and membrane-related GO terms while *y4* (cell-membrane vs. fatty-acid-synthesis) is enriched in membrane-related and transport GO terms. Several nucleotide-binding related, protein-binding, and metal-binding GO terms were enriched in sub-model *y5* in the classification between protein-synthesis and rna-polymerase inhibitors.

As membrane/transport GO terms were expected to be enriched in gene sets that classify MOA targets related to cellular structure and nucleotide/protein binding related terms were expected for gene product synthesis, a multitude of metal ion related GO terms in the classification of protein-synthesis and rna-polymerase inhibitors was not expected. However, this agrees with previous studies that have focused on metal-responsive ECF sigma factors, several of which are activated by iron depletion or by an excess of other metals such as zinc (Moraleda-Muñoz et al., 2019); thus, overlapping with the GO terms enriched in our GSEA analysis. Bacterial ECF sigma factors are directly involved in the transcription process by recognizing promoter sequences, together with the core RNA polymerase enzyme, and initiate the transcription of the genes they regulate (Heimann, 2002). Although the models can be fully understood from a mathematical perspective, biological interpretation is limited to previous empirical studies and the extent of domain knowledge available. However, these methods are expected to provide a powerful resource in guiding empirical validation experiments to demystify complex biological processes.

The CoHEC models present a purely data-driven XAI approach that can predict the primary MOA from unobserved compounds with high performance. This data-driven AI maximizes the available information content by asking simple questions about specific genes in a particular order to effectively evade statistical artifacts that are inherent in biological datasets where features greatly exceed the number of observations.

4.4 - Interactions between fecal gut microbiome, enteric pathogens, and energy regulating hormones among acutely malnourished rural Gambian children

There is increasing evidence that the pathogenesis of acute malnutrition may be multifactorial and linked to microbial dysbiosis, the overgrowth of specific enteric pathogens in the gut and hormonal imbalance. However, most of this evidence has been generated from children with edematous malnutrition in East Africa and Southeast Asia. Despite West Africa being a prominent area for malnutrition, little is known about the pathogenesis of acute malnutrition during childhood in this region; non-edematous being the predominant type of malnutrition in The Gambia. The primary aim **Publication III** was to investigate the role of energy regulating hormones in the variable growth responses of rural Gambian children during nutritional rehabilitation.

This study was conducted in the East, West, and Central Kiang districts of rural Gambia where children from 6 to 24 months of age were recruited by clinicians and given health care services. The clinicians were granted ethical approval to conduct this study by the *Joint Medical Research Council Unit The Gambia* and *Gambia Government Ethics Committee*, L2015.14. With this ethical approval, all children were recruited into the study with written informed parental consent. All participants recruited into the study underwent

clinical assessments and assigned to one for the following groups: moderately acute malnourished (MAM), severely acute malnourished (SAM) SAM, and well-nourished (WN, WHZ > -2) where WHZ refers to weight-for-height Z-score. WN control participants were based on the anthropometric measurements (World Health Organization, 2006). MAM was classified as WHZ between 2 and 3 standard deviations below the WHO growth reference standard or a mid-upper arm circumference between 115 and 125 mm while SAM was classified as WHZ less than 3 standard deviations below or a mid-upper arm circumference less than 115 mm (World Health Organization, 2006). SAM children were managed in an outpatient nutrition rehabilitation unit in rural Gambia with additional health services.

Pre- and 1-hour post prandial venous blood samples for analysis of energy-regulating hormones were collected from all children at recruitment and for children in the MAM and SAM groups at days 14 and 28. Stool samples were collected from all the children at recruitment and at follow up visits at days 14 and 28 for children that presented with MAM or SAM.

The approach of **Publication III** leverages information gained from a sample-specific perspective to provide insight into the defining characteristics of undernutrition at different stages. To investigate the role of energy regulating hormones in the variable growth responses of rural Gambian children during nutritional rehabilitation, 3 modalities were measured: (1) fecal microbiome marker gene survey metagenomics of 16S rRNA sequences (compositional data); (2) TaqMan Array Card assay of enteric pathogens (binary data); and (3) clinical measurements related to immune activation, inflammation,

and energy regulating hormones (continuous). This study provided novel insights into the role of microbial dysbiosis, enteric pathogens, and host energy-regulating hormones in the pathogenesis of acute malnutrition among West African Children while characterizing key differences between severe and moderate non-edematous acute. As with edematous SAM, non-edematous SAM is also characterized by the collapse of a complex system including the gut microbiome, enteric pathogens, and energy regulating hormones.

4.4.1 – Determining the range of fecal microbial diversity of each nutritional status phenotype

The first objective of **Publication III** was to investigate microbial diversity of each nutritional status phenotype to identify defining characteristics. With respect to the fecal microbiome, 388 high quality OTUs were observed across all participants and the number of detected OTUs (richness) ranged from 14 - 167 per sample (**Publication III** – Fig. 1A). The microbial richness of SAM samples was statistically lower than both WN and MAM samples. MAM samples had statistically higher variance compared to WN and SAM (**Publication III** – Fig. 1B). The variance in richness was the highest for WHZ in the range (-3,-1), the entirety of MAM samples and the lower WHZ of WN samples (**Publication III** – Fig. 1C).

Bacterial biomass is inversely proportional to 16S rRNA qPCR *Ct* values and, supporting the trends in microbial richness, differences in *Ct* values between nutritional status phenotypes were observed. MAM samples exhibited the lowest median *Ct* and thus the highest relative biomass on average (22.5 *Ct*), while SAM exhibited the lowest (30.5 *Ct*) with WN in between (25.4 *Ct*) (**Publication III** – Fig. 1D). A bimodal distribution of SAM

Ct values with the lower peak at ~24 *Ct* and the upper at ~35 *Ct* suggested that while SAM samples have similar community composition, in a substantial number of cases the bacterial community had collapsed from a numerical perspective.

4.4.2 – Differential abundance analysis of individual fecal gut microbes, enteric pathogens, and energy regulating hormones

The second objective of **Publication III** was to investigate differential abundance of each modality individually. As each modality had a unique data type (i.e., compositional, binary, and continuous), modality-specific differential abundance analysis was performed. With regards to the compositional data, the *ALDEx2* (Fernandes et al., 2014) was used as it was designed specifically for microbiome compositions. The clinical measurement distributions were non-gaussian and required a non-parametric test similar to the ANOVA (Kruskal-Wallis H-test). The enteric pathogen screening was binary and required a Fisher's exact test where the binary data was converted into detection event ratios for visualization. After correcting for multiple hypotheses, individual features that were enriched or depleted in WN, MAM, and SAM were determined. This directionality of enrichment or depletion was critical for interpreting the sample-specific network analysis in a clinical context as the perturbations were either negative or positive.

The relative abundances of bacterial phyla were very similar across all nutritional status phenotypes with a few notable exceptions (**Publication III** – Fig. 2A). The combination of *Firmicutes* (WN = 46%, MAM = 50%, SAM = 33%), Bacteroidetes (WN = 28%, MAM = 20%, SAM = 17%), and Proteobacteria (WN = 20%, MAM = 23%, SAM = 45%)

constitute more than 90% of the microbial abundance regardless of the participant's nutritional status with respect to visit. No components at any level of taxonomy were differentially abundant among children with MAM relative to WN. However, an enrichment in *Enterobacteriaceae* abundance for SAM (μ = 42%) relative to MAM (μ = 19%) and WN (μ = 18) was observed at the family level. The only differentially abundant OTU was an unclassified *Klebsiella* (*Otu000014*) with an enrichment in SAM (μ = 16%) relative to MAM (μ = 6%) and WN (μ = 3%).

Analysis of clinical measurements pertaining to energy regulating hormones gave insight into which metabolites were proportional to fluctuations in WHZ. Statistically significant trends between IGF-1, leptin, and IGFBP- 3 (the main binding protein of IGF-1) that increased with WHZ and nutritional status were identified (**Publication III** – Fig. 3A). IGF-1 is a key growth regulating hormone in infancy and plays an important role during nutritional recovery in undernourished children. Similarly, leptin is a hormone predominantly made by adipose cells and enterocytes in the small intestine to help regulate energy balance by moderating appetite and intestinal barrier function (Brennan and Mantzoros, 2006) and plays a major role in signaling energy deficit in acute malnutrition (Bouillanne et al., 2007; Prentice, Moore, Collinson, & O'Connell, 2002). Reverse trends were also observed, that is, the levels of sOB-R (and the molar excess of sOB-R:leptin), ghrelin (and its binding protein) increased monotonically as WHZ decreased. These findings support the hypothesis by Stein et al. 2006 that sOB-R is upregulated during starvation to maintain low levels of bioactive leptin and increase its half-life, thus, decreasing energy expenditure and increasing food uptake. Ghrelin is a well-studied hormone produced by enteroendocrine cells of the gastrointestinal, tract with

substantial production in the stomach, (Kojima et al., 1999; Müller et al., 2015) and circulating ghrelin blood levels are often highest when an individual experiences hunger while returning to lower levels after food intake (Cummings et al., 2001; Müller et al., 2015). The clinical measurements of these metabolites were consistent with previous studies and provided a strong foundation for more complex analytical methods in the context of acute malnutrition (e.g., SSPN analysis).

The abundance of 23 enteric pathogen markers was assessed using qPCR. All children had between 1 to 15 pathogen or virulence factors detected in at least one visit (**Publication III** - Fig. 3B). In accordance with microbial richness, MAM was observed to have a greater number of pathogenic markers (8 pathogens) compared to SAM or WN (7 pathogens) nutritional status phenotypes. Several pathogenic markers were differentially prevalent between the nutritional status phenotypes. In particular, *Giardia duodenalis* was significantly less prevalent among children with SAM (23%) than MAM (68%) or WN (70%). In contrast, the prevalence of enteropathogenic *E. coli* with the *bfpA* and *eae* virulence increased with severity of malnutrition (**Publication III** - Fig. 3B). Bundle-forming pilus A (*bfpA*) and intimin adherence protein (*eae*) are genes found on the EAF plasmid and EPEC genome, respectively, and contribute to attachment to epithelial cells, thus, leading to the attaching and effacing phenotype (Blank et al., 2000; Slinger et al., 2017).

4.4.3 – Multimodal sample-specific perturbation network analysis to quantify changes relative to a reference group

The third objective of **Publication III** was to investigate the interactions between the fecal gut microbiome, enteric pathogens, and energy regulating hormones among rural Gambian children with non-edematous SAM during outpatient nutritional rehabilitation and MAM relative to WN children. However, for each participant there were at most 3 time-ordered samples collected (i.e., t=0, t=14, and t=28 days) which are not enough timepoints for reliable temporal networks using existing methodologies. Instead of using temporal network frameworks, this study developed a novel methodology to explore sample-specific networks that can be analyzed with respect to individual timepoints or a trajectory of timepoints.

The objective with a sample-specific approach was to maximize the available data and quantify the amount by which a particular sample perturbs a biologically relevant background distribution; referred to as a sample-specific perturbation network (SSPN). Perturbation in the context of SSPNs is defined as a change in association strength of an edge between background network and perturbed background network distributions. For the background cohort, data was used from participants whose nutritional status was WN for all, which allowed for the calculation of SSPNs for participants that deviated to or from WN between visits. This implementation allows one to quantify how much a given sample can perturb a group of reference samples by either increasing or decreasing association strength between two features. SSPN can be powerful for characterizing undernutrition when used in the context of enrichment or depletion of individual features relative to nutritional status.

4.4.4 – Identifying perturbations capable of discriminating nutritional status phenotypes

Network analysis, especially fully-connected networks, often produces a plethora of edges to investigate making it difficult to interpret the results. Typical association networks use all the samples to build a single network but with the SSPNs developed in this study, there is a separate network for every sample producing a perturbation matrix (N= 82 samples, M= 14,270 edges representing 188 nodes). This data structure produces a unique opportunity that allows one to leverage machine learning algorithms using the perturbation matrix as a feature matrix. This has a strong advantage because one can use a mixture of supervised (e.g., feature selection) and unsupervised (e.g., network analysis) to gain insight into the latent structure of the disease with respect to the 3 modalities used to build the networks.

By analyzing the SSPNs using the HEC model and *Clairvoyance* feature selection from **Publication II** in unison (i.e., CoHEC model), along with the fundamentals established in **Publication I**, the edges that could predict nutritional status based on perturbations were identified (**Publication III** – Fig. 4A). Essentially, this characterizes which interactions are most predictive of nutritional phenotype and is more informative for nutritional intervention than regression methods predicting a rise or fall in WHZ and disregarding intra-phenotype patterns. By selecting for only the edges that are informative in discriminating nutritional status phenotypes, and by extension the informative nodes, the information content in the edges was compressed by 98.143%; that is, 265 of the 14,270 edges. As a collection of edges can be combined into a network, one can reconstruct the structure of the network

with only edges that were relevant for characterizing nutritional status phenotype referred to as an aggregate network (AN) (**Publication III** – Fig. 5).

4.4.5 – Evaluating nutritional status predictive performance on new patients

Unsupervised machine-learning can be used to gain insight into the underlying structure of the data and leveraged to validate dimensionality reduction (e.g. feature selection) results based on predefined categories, nutritional status in this context. To prevent the models from overfitting, a similar approach to LCOCV from **Publication II** was used but instead of leaving individual compounds out, participants were left out to simulate performance on new children. These models were able to predict nutritional status phenotype at 100% accuracy using the perturbations of the sample with respect to a reference group despite the rigorous performance metric. Unsupervised clustering of held-out prediction probabilities were more homogenous than unsupervised clustering based on microbiome abundance profiles or pathogen markers (**Publication III** – Fig. 4C). Clustering by held-out prediction probabilities revealed that the edge features in each sub-model capture biologically relevant discriminatory patterns. As logistic regression classifiers were used under the hood in the CoHEC models, it is possible to completely understand how each edge contributes to the prediction process. Furthermore, the increased homogeneity of nutritional status clustering after running the CoHEC model demonstrates how the XAI was able to learn hidden structure within the data.

4.4.6 – Quantifying changes in sample-specific perturbation networks over time with recovery scores

This study contained time-ordered samples for many of the participants. An edge recovery score was developed in **Publication III** to quantify the amount in which an edge contributes to weight recovery; more specifically, the transition from MAM or SAM to WN. As SSPN edge weight indicates a perturbation relative to a reference group, perturbations relevant to nutritional status recovery were identified by selecting for edges in consecutive time-ordered SSPNs that have the following properties: (1) greatest change in perturbation of associations between visits t_n and t_{n+1}; and (2) smallest perturbation magnitude from the reference group at t_{n+1} where t_n and t_{n+1} phenotypes represent MAM or SAM and WN, respectively.

The recovery score metric condensed the information content of complex multimodal time-ordered SSPNs into a single human interpretable metric. The recovery score was designed to demonstrate the following properties for each edge: (1) a large difference between an undernourished visit and a consecutive WN visit; and (2) a small edge weight for WN (**Publication III** – Fig. 6A). Emphasizing these properties made it possible to collapse the temporal dimension, the sign of edge weights, and focus specifically on edges specifically to the recovery of an individual participant. Although the cohort sample size was not large enough to statistically model specific associations in a participant's ability to decline or recover with respect to their nutritional status these methods will be useful for larger studies in the future.

4.4.7 – *Interpreting CoHEC models and SSPNs in the context of acute malnutrition*

Analyzing each modality independently presented a means to validate each finding by cross-referencing against previous studies, thus, setting the context for multimodal

network analysis. Consistent with previous reports (Alou et al., 2017; Gough et al., 2015; Smith et al., 2013; Subramanian et al., 2014), children with SAM had significant reductions in richness and bacterial biomass compared to WN or MAM participants. It is possible that the bacteria depleted in acutely malnourished children are essential for optimal digestion, nutrient absorption, modulating inflammation and immune development (Smith et al., 2013). A new finding from this study was that children with MAM had statistically significant enrichments in the gut bacterial loads and variance in microbial richness compared to WN and SAM participants. These changes may be an indication of impaired immune function in the children with acute malnutrition (Jones and Berkley, 2014) which agrees with our findings that IGF-1 and leptin are associated with various microbes and are highly predictive of nutritional status.

Enterobacteriaceae abundance was greatly enriched in SAM children and may be linked to the low prevalence of *Giardia* which competes for the same ecological niche in the small intestine (Allain et al., 2017). However, previous research using mouse models showed that *Enterobacteriaceae* was over-represented in *Giardia* infected mice (Bartelt et al., 2017). There appears to be a more complex mechanism regulating the balance between *Enterobacteriaceae* (bacteria) and Giardia (protozoan parasite). These finding may therefore be specific for masasmic SAM involving the differential regulation of anti-parasitic and anti-bacterial immune responses in these children where *Giardia* infection alters immune responses to *E.coli* or vice versa. This warrants further exploration. In addition, EPEC virulence factors *bfpA* and *eae* were observed to increase with increasing severity of malnutrition. EPEC adheres to intestinal epithelial cells, causing diarrhea, and

constitutes a significant risk to health, especially in very young children (Chen and Frankel, 2005). Subramanian and colleagues also reported an enrichment of *Enterobacteriaceae spp.* among children with SAM from Bangladesh (Platts-Mills et al., 2017; Subramanian et al., 2014); although, a causal pathway is yet to be identified.

This study found that *Escherichia/Shigella* sp. and molar ratio of sOB-R:letpin had substantially greater predictive capacity in discriminating WN from undernourished participant samples compared to other nodes. In particular, *Escherichia-Shigella* had high predictive capacity through its associations with *ghrelin* and *ghrp*. This was not surprising as leptin is a key player in regulating both antimicrobial peptides and microbiota composition and as such, *Escherichia-Shigella* and molar-excess soluble leptin may play pivotal roles in mediating complex interactions that modulate nutritional status. High predictive capacity of molar excess of sOB-R:leptin through *Lactobacillus mucosae,* an unclassified *Haemophilus* and an unclassified *Ruminococcaceae* UCG-002 was also observed. Previous research has identified strong associations between leptin and *Lactobacillus* and it is believed that leptin can modulate gut microbiota by stimulating mucin production which may favor bacterial growth (El Homsi et al., 2007). *Lactobacillus* has been shown to maintain intestinal homeostasis and is speculated to attenuate the pro-inflammatory signaling induced by *Shigella* after invasion of epithelial lining (Tien et al., 2006). Similarly, previous research has ascertained that leptin supplementation resulted in a higher proportion of *Ruminococcaceae* (Grases-Pintó et al., 2019).

Another intriguing finding was the high predictive capacity of perturbations in *IGF-1* and an unclassified *Enterobacteriacea* in discriminating MAM from SAM. The high predictive capacity of perturbations in *IGF-1* and the *Enterobacteriaceae* associations are relevant as *Enterobacteriaceae* are often enriched in children who are wasted along with decreased plasma IGF-1 concentrations (Bartz et al., 2014) and decreased concentrations of IGF-1 and IGFBP- 3 have been observed in underweight mice (Schwarzer et al., 2016). *Enterobacteriaceae* are often be enriched in undernourished individuals (Million et al., 2017) and coupled with decreased concentrations of IGF-1 (Bartz et al., 2014) and IGFBP- 3 (Schwarzer et al., 2016). These findings from other research groups agree with our results showing that IGF-1 and IGFBP- 3 concentrations decrease with WHZ and are lowest in SAM. As immunity is heavily impaired in children experiencing SAM (Hossain et al., 2015), the predictive associations between *Enterobacteriacea* and IGF-1 are not surprising. However, it is not uncommon for children experiencing SAM to develop septicaemia (Hossain et al., 2015; Jones and Berkley, 2014). Previous research has shown that patients with sepsis have low levels of IGF-1 inversely correlated with enteric bacterial load (Hunninghake et al., 2010). Hunninghake et al. 2010 also supposed that translocation of bacteria across the gastrointestinal tract may occur.

Acute malnutrition is a complex multifactorial disease with interplay between the gut microbiome, energy regulating hormones, and the presence of enteric pathogens. It appears that WN systems are stable but as a child's weight declines, approaching MAM, the community destabilizes with increased microbial diversity and interactions. As a

child's nutritional status deteriorates the gut microbiota community becomes depleted and dominated by pathogenic *Enterobacteriaceae* in an ecological collapse as demonstrated by low bacterial load, and low microbial diversity.

4.5 - Differential network analysis of oral microbiome metatranscriptomes identifies community scale metabolic restructuring in dental caries

Dental caries is a microbial disease and the most common chronic health condition, especially in adolescence, affecting nearly 3.5 billion people worldwide. Multi-omics are typically used to investigate microbial diseases, but these approaches present significant challenges when balancing biological accuracy and compositional sparsity. The aim of **Publication IV** was to address these challenges and characterize community-scale metabolic interactions that are diagnostic of caries status using a syngery of metagenomics and metatranscriptomics.

This study investigated the supragingival plaque oral microbiome of 91 Australian children while characterizing 658 bacterial and 189 viral metagenome assembled genomes. The challenges in balancing biological accuracy and compositional sparsity were addressed by developing a reproducible pipeline for clustering sample-specific genomes to integrate both metagenomics and metatranscriptomics analysis regardless of biosample overlap. Furthermore, novel feature engineering and compositionally-valid ensemble network frameworks were developed and their utility for investigating regime shifts associated with caries-related dysbiosis was explored.

The analytical methodologies presented in **Publication IV** are useful for characterizing differential community structure was demonstrated. Further, these methods can be applied when hypothesis testing for differential abundance do not capture statistical enrichments or the results from such analysis are not adequate for providing deeper insight into disease. By moving beyond abundances of genomic features to interactions amongst features this study was able to deconvolute the oral communities to identify not only which organisms and metabolic pathways were central to a system but also how these systems were rewired between caries and caries-free systems. These findings provide evidence of a core oral microbiome composed of both bacteria and viruses that were transcriptionally active in all participants regardless of phenotype and increased network complexity in caries-related dysbiosis.

Finally, this study provided evidence that certain organisms shift their carbohydrate metabolism and serve a bridge between phenotypes. The evidence in **Publication IV** supports the hypothesis that caries is a multifactorial ecological disease. This research demonstrated how investigating microbiomes from different vantage points can provide insight into microbial ecosystems and their relevance in health and disease. The techniques developed in this study were designed for generalizability not limited to microbiome research and provided open-sourced implementations via the updated *EnsembleNetworkX* Python package debuted in **Publication III.**

4.5.1 – Isolating individual bacterial and viral genomes in silico

The first objective of this research was to isolate individual genomes from metagenomic assembles (i.e., MAGs). Metagenomes from the Australian children in this study were

evaluated and analyzed previously (Espinoza et al., 2018; Shaiber and Eren, 2019) but substantial improvements in assembly, binning, and quality assessment methodologies warranted revisitation and reanalysis. There exists a plethora of binning algorithms to date, each with their own strengths and weaknesses. To leverage the strengths of different techniques, 4 different types of algorithms were used including *MaxBin2* (Wu et al., 2016), *Metabat2* (Kang et al., 2019), *VAMB* (Nissen et al., 2021), and *VAMB* in multi-binning mode. The resulting bins from these tools were consolidated using a separate algorithm called *DAS Tool* (Sieber et al., 2018) where MAGs were evaluated and refined to produce high quality consensus MAGs from multiple inputs. The final MAGs were filtered using the quality assessment capabilities of *CheckM* (Parks et al., 2015). Said methodology is designed specifically for prokaryotes but preliminary analysis suggests there is a lush community of viruses in the oral microbiome. To isolate viral MAGs, the unbinned contigs were extracted using *VirSorter2* (Guo et al., 2021) and *CheckV* (Nayfach et al., 2020). Since the dataset consists of both metagenomics and metatranscriptomics, both DNA and RNA viruses could be assessed. This consensus binning approach resulted in 658 bacterial, 179 DNA viral, and 10 RNA viral MAGs that could be used as a reference for metatransciptomics.

These bacterial MAGs clustered into 135 unique SLCs representing 49 hitherto unclassified species with 26 of which classified as *Patescibacteria* candidate phyla radiation (CPR; 6 *Gracilibacteria/SR1*, 43 *Saccharibacteria*) with a total of 69 CPR MAGs collectively. Of the non-CPR SLCs, 31 *Bacteroidota,* 22 *Proteobacteria*, 21 *Actinobacteriota*, 23 *Firmicutes(A/C),* 8 *Fusobacteriota,* and 4 *Campylobacterota* were

identified (**Publication IV** - Table 2,S2,S3). The DNA and RNA viruses clustered into 137 and 5 unique SLCs, respectively. Most of the DNA viruses were classified as *Caudovirales,* of unknown species, associated with the human oral (42 SLCs), gut (41 SLCs including 1 *Inoviridae*), human respiratory (1 SLC), and non-specific environments (1 SLC). Aside from these unknown species, several *Caudovirales* phages for *Arthrobacter* (7 SLCs), *Streptococcus* (4 SLCs), *Klebsiella*, *Haemophilus*, *Pasteurella*, *Pseudomonas*, and *Burkholderia* were also identified. Other than *Caudovirales*, *Streptococcus* satellite phages (2 SLCs), unclassified CRESS-DNA *Parvovirus* associated with the human gut (2 SLCs), and an unclassified virus associated were identified within the human oral environment. Most of the RNA viruses were *Escherichia* phages (4 SLCs) designated as Qbeta BZ1, MS2, and BZ13 strains but a novel virus with no close taxonomic classification was also uncovered.

4.5.2 – Species-level clusters and species-specific ortholog analysis

The second objective of this research was to develop a method that could balance biological accuracy with compositional sparsity. There are two main approaches when conducting metagenomic assembly of *N* samples: 1) a consensus assembly where reads from all *N* samples are merged and assembled together for a single assembly; and 2) assembly of each sample individually to produce *N* assemblies. There are pros and cons with each method where option 1) addresses the compositional sparsity issue and 2) addresses the biological accuracy issue. In a consensus assembly all the samples are merged together so the resulting contigs will be a mixture of communities yielding suitable consensus MAGs but ones that are not biologically accurate with high likelihood of

chimeric sequences. Further, the assembly algorithms using a consensus approach take much longer, use more compute resources, and produce lower quality assemblies with fewer longer contigs than if each sample was assembled individually. The benefit of this approach is that the reads can be mapped and the biological features (e.g., contigs, ORFs, MAGs) can be compared directly across samples since all the samples are used to construct these assemblies. For the sample-specific approach, assembling each sample individually allows for quicker processing time, increased parallel processing (as each sample can be assembled separately), and the resulting biological features are more biologically accurate since there are fewer genomic variants in 1 sample compared to *N* samples. The caveat of this approach is the biological features generated are sample-specific and not directly comparable across samples.

The solution in **Publication IV** to "have your cake and eat it at the same time" that addressed this limitation was to develop a hierarchical structure for biological features that could be collapsed and expanded *in silico*. More specifically, the MAGs were clustered into species-level clusters (SLC) using 95% average nucleotide identity (ANI) via *FastANI* (Jain et al., 2018) (**Publication IV** – Fig. 1). This allowed for sample-specific strains to retain their biological accuracy (instead of merging assemblies for chimeric contigs) while providing a means to collapse the MAGs into SLCs that can be compared across samples. Though, this approach does not make it possible to analyze specific genes across different samples.

To overcome the challenge of comparing genes across different samples, the genes for each SLC were clustered via ortholog analysis to yield SLC-specific orthogroups using *OrthoFinder* (Emms and Kelly, 2019). This presented a unique opportunity to collapse the same genes from different MAGs within a SLC for direct comparisons between samples. This effectively addressed the inherent sparsity issue that stymied the path towards emphasizing biological accuracy over analytical practicality. That is, this approach is biologically accurate and not inherently sparse. This research addresses a critical limitation in paired metagenomics and metatranscriptomics: that is, how to have biologically accurate assemblies not biased by coassembled chimeric contigs while also producing overlapping features (e.g., SLC, SLC-specific orthogroups). Lastly, this reproducible methodology is applicable for all domains of life and not limited to this study (detailed in **Publication IV** – Fig. S1).

4.5.3 – A core oral microbiome of bacteria and viruses

This study provided evidence of a core bacterial and viral oral microbiome across this cohort of Australian children regardless of collection center, age, or sex. The core bacterial microbiome exists at the genus level as almost every genus is transcriptionally active in every sample (Clusters 2.1-2.4 from **Publication IV** - Fig. 2), regardless of phenotype, but this was not the case for most viruses. Though, there were 2 DNA viruses that were highly prevalent across samples that could be considered part of the core oral microbiome.

This core microbiome supports the ecological plaque hypothesis that environmental conditions influence the metabolism of existing microbes nudging the community into a cariogenic configuration. As the oral community is able to shift the collective metabolism to adapt to a cariogenic environment, the reverse must also be true given the prevalence of this core community in both conditions. The implications of such a finding propose the possibility for diagnostic therapeutics for caries detection and caries prevention via probiotics. Characterizing the interactions between microbes and their additive metabolism is expected to provide a deeper insight into what it means metabolically to have a cariogenic oral microenvironment and, also important, a caries-free microenvironment.

4.5.4 – Microbiome feature engineering to couple taxonomy with functionality

Within this dataset, there were 255,737 SLC-specific orthogroups which would result in ~32.7 billion non-redundant connections in a fully-connected coexpression network; an insurmountable dataset for exploratory analysis on most modern machines. Instead of using draconian filtering thresholds of orthogroups, this was addressed by the third objective for this study which was a feature engineering technique that would allow seamless transitions from read ↔ ORF/orthogroup ↔ contig/MAG/SLC ↔ engineered feature using custom taxonomy fields and functional assignments (e.g. KEGG, MetaCyc, PFAM).

Publication IV introduced the PhyloGenomic Functional Category (PGFC) as a supervised microbiome feature engineering method for high-level statistical analysis that

could be expanded back into the underlying orthogroups (or ORFs); unlike dimensionality reduction methods such as PC[o]A, *t*-SNE (Van Der Maaten, 2014), or UMAP (McInnes et al., 2018). PGFCs essentially group low-level features, orthogroups in this context, by a taxonomic unit (SLC) and a functional unit (KEGG module) (**Publication IV** - Fig. 3B, Table S5); similar, but not identical, to *HUMAnN* (Beghini et al., 2021; Franzosa et al., 2018) which does not allow the flexibility for custom low-level features from *de-novo* meta-omics. Another similar approach is the *amalgram* where compositions can have either exclusive or non-exclusive mappings between the original feature and engineered feature (Quinn and Erb, 2020). However, these engineered features cannot be collapsed and expanded with respect to predefined categories such as taxonomy and metabolism so will not be explored in this study.

PGFCs are composite features that group metabolic functional information with genome-resolved taxonomy assignments and were created by grouping all of the orthogroups that had KEGG orthology, defined via KOFAMSCAN (Aramaki et al., 2020), and extending the grouping up the hierarchy to modules with respect to taxonomy. Taxonomy for PGFCs was assigned to the SLC of origin. PGFCs were implemented using the *EnsembleNetworkX* Python package updated with **Publication IV** methodologies.

The quality assessed PGFC dataset contained 2,478 PGFCs representing 89 taxonomic units and 113 functional units from 8,554 orthogroups; all of which are from bacterial SLCs. In orthogroup-space, these features would amount to ~37 million non-redundant connections but only ~3 million% in PGFC-space, effectively compressing the information

content by ~92%, making prototyping and data exploration tractable on modern compute machines.

4.5.5 – Characterizing metabolic structures unique to each phenotype

The fourth objective of this study was to investigate the unique and shared characteristics of caries and caries-free microbial community metabolism. Network theory is an advantageous framework to pursue this type of systems biology as the associations between biological features can be interpreted as (indirect) interactions within the environment. Many traditional approaches use all the available data to construct a single network and then compare the differential abundance of individual features between two conditions. While this may be suitable for characterizing high-level mechanisms, this approach does not fully harness the potential of network analysis; that is, the connectivity between biological features. To understand how biological interactions change between phenotypes, a combination of ensemble networks (*EnsembleNetworkX* debuted in **Publication III**) and compositionally-valid phenotype-specific coexpression networks (PSCN) were implemented producing robust networks for caries and caries-free phenotypes separately. Further, these networks use PGFCs as nodes so that taxonomy and functionality can be coupled during network inference and interpretation. This approach allowed not only for characterization of each phenotype individually but also for comparison between phenotypes through inferred interactions; the true structure of the network is unknown *a priori*.

Neisseria appears to be a key player with high connectivity in the supragingival plaque oral microbiome regardless of caries phenotype. Previous research has observed *Neisseria* as highly abundant in both caries and caries-free microbiomes (Yang et al., 2021) but this study was the first to report this trend in the context of network connectivity. Although the connectivity of *Neisseria* is comparable in both microbiomes, the high connectivity in the caries microbiome is masked by a plethora of other highly connected genera and is ranked higher in the caries-free microbiome as a result of fewer high connectivity genera (**Publication IV** - Fig. 4A,B). However, different microbial communities were observed interacting with *Neisseria* when comparing between caries and caries-free microbiomes. In particular, several species of *Neisseria* were interacting with members of *Bacteroidota* in the caries-free microbiome and shifts to interactions with *Haemophilus D parainfluenzae* and fellow *Neisseria* in the caries microbiome (**Publication IV** - Fig. 4E). This is interesting because several species of *Neisseria* had enriched connectivity in the caries-free microbiome and *Haemophilus D parainfluenzae* had enriched connectivity in the caries microbiome (**Publication IV** - Fig. 5A,C). Although, *Neisseria* and *Haemophilus parainfluenzae* are both common in the oral cavity of caries-free individuals from the perspective of abundance (Keijser et al., 2008; Liljemark et al., 1984; Zaura et al., 2009), their interactions with other coexpressed microbes known to be associated with infections in humans may be indicative of caries dysbiosis.

4.5.6 – Characterizing community scale metabolic restructuring using differential coexpression networks

Differential coexpression networks (DCN) can reveal changes in connectivity between a reference and treatment network. As ensemble PSCNs are the building blocks of DCNs in this study, the DCNs provided the same benefits with respect to compositional validity and outlier resistance. Previous approaches have used DCNs but did not use compositionally-valid association metrics nor ensemble networks (Fuller et al., 2007; Hsu et al., 2015). While differential abundance/expression analyses can be useful in identifying feature enrichment (e.g., OTU, MAG, ORF, gene, etc.), each method has their own caveats in assumptions about the data distributions (well characterized in (Morton et al., 2019) with the establishment of reference frames) and provide no information regarding differences in pairwise interactions; an essential perspective when studying diseases resulting from dysbiosis. Using the $PSCN_{Caries\ free}$ as a reference network and $PSCN_{Caries}$ as the treatment network, this study was able to construct a DCN from statistically significant PGFCs, determined from PSCN analysis, for seamless cross-referencing between PSCNs and the DCN.

Unsupervised clustering of the DCN revealed 6 clusters (**Publication IV** - Fig. 6, Table S5,6), of which there were 3 high connectivity DCN clusters (HCDC), each being diagnostic of phenotype. For the only cluster with connectivity enriched in the caries microbiome, the differential connectivity was primarily from *Capnocytophaga sputigena*, *Kingella B oralis, Vellonella parvula A*, *Streptococcus sanguinis*, *Streptococcus oralis*, and unclassified *Streptococcus* via carbohydrate and cofactor/vitamin metabolism (**Publication** IV - Fig. S4).

The ability to collapse and expand PGFCs in these abstract network spaces can be used to identify unanticipated players with uncharacterized interactions relevant to maintaining either caries-free or caries microbiomes. For instance, *Cardiobacterium hominis* emerged as a hub not only in the caries-free microbiome but also in the caries microbiome primarily through ATP synthesis and carbohydrate metabolism. *Cardiobacterium hominis* was a link between the highest differential connectivity clusters for caries and caries-free microbiomes through carbohydrate metabolism and ATP synthesis. The most unexpected finding was that *Cardiobacterium hominis* citrate cycle and fumarate reductase were highly centralized suggesting a shift in carbohydrate metabolism from pentose phosphate cycle to citrate acid cycle in the caries microbiome.

4.5.7 – Interpreting PSCNs and DCNs in the context of dental caries

The ensemble networks implemented in this study implicated *Cardiobacterium hominis* as a nexus between caries-free and caries dysbiotic states through a transition from pentose phosphate to TCA cycle carbohydrate metabolism. Previous metabolic research confirms that both the TCA cycle and the pentose phosphate pathway function within the supragingival plaque *in vivo* and glycolytic activation causes an increase in pentose phosphate activity (Takahashi et al., 2010). These findings suggest that *Cardiobacterium hominis* mediated pentose phosphate pathway metabolism promotes a caries-free microbiome with the support of *Streptococcus sanguinis* lysine metabolism, *Abiotrophia sp001815873* ATP synthesis*,* and *Neisseria* cofactor metabolism (Community-7.II from **Publication IV** – Fig. 7*).* This hypothesis agrees with previous research as *Streptococcus sanguinis* and *Abiotrophia* have been known to cooccur in caries-free children (Kanasi et

al., 2010) while *Neisseria,* as mentioned previously, has been associated with beneficial oral health. The simplicity of interactions enriched in the caries-free microbiome agreed with this study's finding that fewer taxa with more defined metabolisms are indicative of stable and healthy oral communities; thus, opening the door for potential probiotics, engineered microbial communities, and therapeutics for oral health and resilience.

The evidence for *Cardiobacterium hominis* TCA cycle and its association with caries dysbiosis was more complex as it had considerably more taxa and metabolic pathways than communities including pentose phosphate pathway. However, this agreed with the finding that caries-related regime shifts include more high connectivity interactions without an increase in microbial richness; that is, greater total connectivity with the same core microbiome. Previous research has shown that the caries microbiome has the potential to metabolize more diverse sugar source than the caries-free microbiome (Espinoza et al., 2018) which supports the notion that caries dysbiosis has more complex metabolism and, therefore, higher total network connectivity. *Cardiobacterium hominis* TCA cycle had enriched connectivity to carbohydrate metabolism from *Kingella oralis* (Cherkasov et al., 2019), *Streptococcus oralis* (Kanasi et al., 2010), and *Corynebacterium matruchotii* (Yang et al., 2021) in the caries microbiome which have previously been statistically associated with caries dysbiosis in children. In the DCN community detection analysis, *Cardiobacterium hominis* was the only microbe that had connectivities enriched in caries and in caries-free microbiomes which supports the hypothesis of turncoat behavior with respect to oral health.

5 – CONCLUSION

The major conclusions based on the objectives of the thesis are as follows:

- **Publication I**
    - NGS data is compositional, this is not optional, and compositionality must be incorporated when samples are being compared.
    - When using relative abundances, the distance between variables is sensitive to the presence or absence of individual components and can reveal spurious relationships amongst unrelated variables resulting in false positive correlations.
    - Aitchison geometry addresses the compositionality of NGS data and provides a basis for many transforms that accurately represent NGS data.
    - Log-ratio transformations, a form of Aitchison geometry, perform equivalently on both the counts and proportions while capturing the relationships between features within the sample space.
    - Proportionality is a compositionally-valid association metric that can roughly be included as a drop-in replacement for correlation.
    - Association networks can provide a basis for more complex network structures such as DCNs and SSPNs.
- **Publication II**
    - Based on *E. coli* transcriptional responses from compounds of the same MOA suggested that established MOA categories are a spectrum of biological responses rather than discrete entities.

- Feature selection used synergistically with hierarchical classification models (i.e., CoHEC) are effective in classifying complex categories within noisy datasets.
- CoHEC models can accurately predict the MOA of transcriptomic responses challenged with crude extract when trained on pure compounds.
- CoHEC models can exceed the performance of a multiclass model using the same base algorithm (e.g., logistic regression in this study) with only a fraction of the training data when using the same input features.
- These probabilities computed by the CoHEC model on test sets are machine informative as unsupervised analysis of these probabilities clusters compounds by MOA with statistically greater homogeneity than the input data.
- CoHEC models can flag compounds with novel activity and classes not represented by the training data as was demonstrated with darobactin.
- CoHEC models are generalizable to non-compositional data such as metabolomics and perform highly in other biological systems (e.g., *Mycobacterium smegmatis* and other strains of *E. coli*)

- **Publication III**
  - Children experiencing SAM have lower microbial richness than MAM and WN.
  - Children experiencing MAM have greater variability in microbiome compositional compared to SAM and WN.
  - Bacterial biomass overall is lower in SAM children.

- The MAM phenotype appears to be a point of instability between two “stable” microbiome configurations where stability refers to low variance communities; that is, WN and SAM.
- Ensemble network can be used as a basis for precision medicine techniques via SSPN and can be adapted for multimodal datasets.
- Feature selection can be repurposed as a community detection algorithm when applied to SSPNs allowing edges relevant to a particular phenotype to be identified.
- *Giardia* was inversely proportional to enteropathogenic *E. coli* virulence in children experiencing SAM suggesting competition for the same ecological niche.
- Unique associations between sOB-R:leptin with *Lactobacillus mucosae*, an unclassified *Haemophilus* and an unclassified *Ruminococcaceae* UCG-002 were identified to be predictive of nutritional status.

- **Publication IV**
  - There exists a core oral microbiome of bacteria and viruses in the cohort of Australian children and this prevalence is maintained despite differences in geographic region, sex, and age.
  - The balance between biological accuracy and compositional sparsity can be addressed by assembling samples individually to obtain sample-specific MAGs, retaining their biological accuracy with minimal chimeric contigs, and clustering MAGs into SLCs that can be compared across non-overlapping samples.

- PGFCs can be used for compositionally-valid microbiome feature engineering to analyze large meta-omics datasets while coupling taxonomy to functionality.
- Ensemble networks can be applied using compositionally-valid association metrics and can provide the basis for PSCNs and DCNs.
- Many species of *Neisseria* interacted with members of *Bacteroidota* in the caries-free microbiome while shifting to interactions with *Haemophilus D parainfluenzae* and fellow *Neisseria* in the caries microbiome.
- *Cardiobacterium hominis* appears to be a key species in the transition from a caries-free to caries microbiome via a switch in pentose phosphate to TCA cycle metabolism.

6 – RECOMMENDATIONS

Based on the findings from this research, recommendations for future studies on characterizing microbial-related diseases using XAI from a systems perspective are as follows:

- Developing benchmarks and exploring how the sampling size and number of iterations effects the distributions in the ensemble networks will be critical for optimizing their usage across different types of datasets.
- Adapting the *Clairvoyance* feature selection algorithm to handle regression scenarios and making it more scalable using neural network frameworks would drastically improve the generalizability.

- Benchmarking the findings of DCNs and SSPNs on the same dataset against ground truths would allow for a rubric for which framework is the most applicable to a particular research question.
- Adapting the SLC pipeline into a single executable command line program that handles eukaryota in addition to bacteria and viruses will be critical for generalizability to metagenomics datasets from any source.
- Further explore the normalization of connectivity measurements and how these influence findings when comparing PSCNs.

7 – EPILOGUE

Developing methodologies through the lens of XAI promotes generalizability for widespread usage and integrates interpretability into the design so the resulting models not only have high performance but also provide insight into the underlying mechanisms; a key feature for biotechnology applications. The development of open-source and reproducible analysis software coupled with the principles of XAI has the potential of bringing on a golden age in characterizing microbial-related diseases.

Despite the progressive techniques recently developed to interpret biological systems, application of XAI in the nascent field of systems biology is far from the status of omniscient. Not knowing the true topology of a system *a priori* inherently limits our approaches towards fully understanding a system's natural complexity. Furthermore, biological systems are not static and modeling the transition between states will yield more intuitive insights on the schematics of these complex structures. The aphorism that "all models are wrong, but some are useful" (Box et al., 2009) holds truth in the paradigm

of inference-based systems biology where knowing the true network structure of an abstract space *a priori* is not attainable. Part of the reason why biological networks are complex is because they utilize abstract constructs (e.g., gene as nodes and proportionality as edge weight) to model observed phenomenon such as disease. This complexity is the aftermath of the uncertainty of true associations, the sensitivity of the methods to infer associations, unaccounted variance (e.g. unknown phenotype), and the dynamics of how these abstractions evolve over time. The abstract space defined by a network is the source of its versatility while also representing the crux of germane interpretation.

There exists great potential for the combined efforts of XAI and network analysis to address the world's most pressing issues. Imagine the synergy of XAI, system-wide cellular modeling (Ebrahim et al., 2013), and "network-of-networks" (multi-level network) frameworks (Gao et al.) harnessed by domain experts spanning climate science to microbiology, public health to agriculture, and from economics to politics modeling the complex flux of resources; an interdisciplinary effort to usurp climate change by identifying solution states that are not only environmentally sustainable but economically productive.

The future of XAI and systems biology must be approached from creative vantage points by building combinatorically on the cornerstones of established concepts, understanding the assumptions of various statistical methods, and interpreting these mathematical abstractions in the context of insightful biological questions where domain knowledge is of utmost importance. The combination of domain expertise, advanced analytical

methods, and creative minds is the foundation of cutting-edge science. Modeling complex systems has provided insight in the past and will certainly continue to do so in the future with the evolution of network theory and the inventiveness catalyzed by the human mind and machines to decipher latent patterns embedded within natural and abstract systems.

8 - REFERENCES

Aitchison, J., 1982. The Statistical Analysis of Compositional Data. J. R. Stat. Soc. Ser. B 44, 139–160. https://doi.org/10.1111/j.2517-6161.1982.tb01195.x

Allain, T., Amat, C.B., Motta, J.P., Manko, A., Buret, A.G., 2017. Interactions of Giardia sp. with the intestinal barrier: Epithelium, mucus, and microbiota. Tissue Barriers. https://doi.org/10.1080/21688370.2016.1274354

Alneberg, J., Bjarnason, B.S., De Bruijn, I., Schirmer, M., Quick, J., Ijaz, U.Z., Lahti, L., Loman, N.J., Andersson, A.F., Quince, C., 2014. Binning metagenomic contigs by coverage and composition. Nat. Methods 11, 1144–1146. https://doi.org/10.1038/NMETH.3103

Alou, M.T., Million, M., Traore, S.I., Mouelhi, D., Khelaifia, S., Bachar, D., Caputo, A., Delerce, J., Brah, S., Alhousseini, D., Sokhna, C., Robert, C., Diallo, B.A., Diallo, A., Parola, P., Golden, M., Lagier, J.C., Raoult, D., 2017. Gut bacteria missing in severe acute malnutrition, can we identify potential probiotics by culturomics? Front. Microbiol. 8. https://doi.org/10.3389/fmicb.2017.00899

Altman, N., Krzywinski, M., 2018. The curse(s) of dimensionality. Nat. Methods 15, 399–400. https://doi.org/10.1038/S41592-018-0019-X

Antipov, D., Hartwick, N., Shen, M., Raiko, M., Lapidus, A., Pevzner, P.A., 2016.

plasmidSPAdes: assembling plasmids from whole genome sequencing data. Bioinformatics 32, 3380–3387. https://doi.org/10.1093/BIOINFORMATICS/BTW493

Antipov, D., Raiko, M., Lapidus, A., Pevzner, P.A., 2020. Metaviral SPAdes: assembly of viruses from metagenomic data. Bioinformatics 36, 4126–4129. https://doi.org/10.1093/bioinformatics/btaa490

Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., Ogata, H., 2020. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. Bioinformatics 36, 2251–2252.

Bartelt, L.A., Bolick, D.T., Mayneris-Perxachs, J., Kolling, G.L., Medlock, G.L., Zaenker, E.I., Donowitz, J., Thomas-Beckett, R.V., Rogala, A., Carroll, I.M., Singer, S.M., Papin, J., Swann, J.R., Guerrant, R.L., 2017. Cross-modulation of pathogen-specific pathways enhances malnutrition during enteric co-infection with Giardia lamblia and enteroaggregative Escherichia coli. PLOS Pathog. 13, e1006471. https://doi.org/10.1371/JOURNAL.PPAT.1006471

Bartz, S., Mody, A., Hornik, C., Bain, J., Muehlbauer, M., Kiyimba, T., Kiboneka, E., Stevens, R., Bartlett, J., St Peter, J. V., Newgard, C.B., Freemark, M., 2014. Severe acute malnutrition in childhood: Hormonal and metabolic status at presentation, response to treatment, and predictors of mortality. J. Clin. Endocrinol. Metab. 99, 2128–2137. https://doi.org/10.1210/jc.2013-4018

Bashiardes S., Zilberman-Schapira, Elinav, E., 2016. Use of Metatranscriptomics in Microbiome Research. Bioinform. Biol. Insights 10, 19–25. https://doi.org/10.4137/BBI.S34610

Başţanlar, Y., Özuysal, M., 2014. Introduction to Machine Learning, in: Yousef, M.,

Allmer, J. (Eds.), MiRNomics: MicroRNA Biology and Computational Analysis. Humana Press, Totowa, NJ, pp. 105–128. https://doi.org/10.1007/978-1-62703-748-8 7

Bastolla, U., Fortuna, M.A., Pascual-García, A., Ferrera, A., Luque, B., Bascompte, J., 2009. The architecture of mutualistic networks minimizes competition and increases biodiversity. Nature 458, 1018–1020. https://doi.org/10.1038/NATURE07950

Beghini, F., McIver, L.J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., Mailyan, A., Manghi, P., Scholz, M., Thomas, A.M., Valles-Colomer, M., Weingart, G., Zhang, Y., Zolfo, M., Huttenhower, C., Franzosa, E.A., Segata, N., 2021. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. Elife 10. https://doi.org/10.7554/ELIFE.65088

Bellman, R., 2003. Dynamic programming. Dover Publications.

Bian, G., Gloor, G.B., Gong, A., Jia, C., Zhang, W., Hu, J., Zhang, H., Zhang, Y., Zhou, Z., Zhang, J., Burton, J.P., Reid, G., Xiao, Y., Zeng, Q., Yang, K., Li, J., 2017. The Gut Microbiota of Healthy Aged Chinese Is Similar to That of the Healthy Young. mSphere 2. https://doi.org/10.1128/msphere.00327-17

Blank, T.E., Zhong, H., Bell, A.L., Whittam, T.S., Donnenberg, M.S., 2000. Molecular variation among type IV pilin (bfpA) genes from diverse enteropathogenic Escherichia coli strains. Infect. Immun. 68, 7028–7038. https://doi.org/10.1128/IAI.68.12.7028-7038.2000

Bowers, R.M., Kyrpides, N.C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T.B.K., Schulz, F., Jarett, J., Rivers, A.R., Eloe-Fadrosh, E.A., Tringe, S.G.,

Ivanova, N.N., Copeland, A., Clum, A., Becraft, E.D., Malmstrom, R.R., Birren, B., Podar, M., Bork, P., Weinstock, G.M., Garrity, G.M., Dodsworth, J.A., Yooseph, S., Sutton, G., Glöckner, F.O., Gilbert, J.A., Nelson, W.C., Hallam, S.J., Jungbluth, S.P., Ettema, T.J.G., Tighe, S., Konstantinidis, K.T., Liu, W.-T., Baker, B.J., Rattei, T., Eisen, J.A., Hedlund, B., McMahon, K.D., Fierer, N., Knight, R., Finn, R., Cochrane, G., Karsch-Mizrachi, I., Tyson, G.W., Rinke, C., Lapidus, A., Meyer, F., Yilmaz, P., Parks, D.H., Eren, A.M., Schriml, L., Banfield, J.F., Hugenholtz, P., Woyke, T., 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat. Biotechnol. 2017 358 35, 725–731. https://doi.org/10.1038/nbt.3893

Box, G.E.P., Luceño, A., Paniagua-Quiñones, M.D.C., 2009. Statistical Control by Monitoring and Adjustment. John Wiley & Sons, Inc., Hoboken, NJ, USA. https://doi.org/10.1002/9781118164532

Brennan, A.M., Mantzoros, C.S., 2006. Drug Insight: The role of leptin in human physiology and pathophysiology - Emerging clinical applications. Nat. Clin. Pract. Endocrinol. Metab. https://doi.org/10.1038/ncpendmet0196

Burchill, E., Lymberopoulos, E., Menozzi, E., Budhdeo, S., McIlroy, J.R., Macnaughtan, J., Sharma, N., 2021. The Unique Impact of COVID-19 on Human Gut Microbiome Research. Front. Med. 0, 267. https://doi.org/10.3389/FMED.2021.652464

Bushmanova, E., Antipov, D., Lapidus, A., Prjibelski, A.D., 2019. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. Gigascience 8, 1–13. https://doi.org/10.1093/GIGASCIENCE/GIZ100

Chaumeil, P.A., Mussig, A.J., Hugenholtz, P., Parks, D.H., 2020. GTDB-Tk: a toolkit to

classify genomes with the Genome Taxonomy Database. Bioinformatics 36, 1925–1927. https://doi.org/10.1093/BIOINFORMATICS/BTZ848

Chen, H.D., Frankel, G., 2005. Enteropathogenic Escherichia coli: Unravelling pathogenesis. FEMS Microbiol. Rev. https://doi.org/10.1016/j.femsre.2004.07.002

Cherkasov, S. V., Popova, L.Y., Vivtanenko, T. V., Demina, R.R., Khlopko, Y.A., Balkin, A.S., Plotnikov, A.O., 2019. Oral microbiomes in children with asthma and dental caries. Oral Dis. 25, 898–910. https://doi.org/10.1111/ODI.13020

Compeau, P.E.C., Pevzner, P.A., Tesler, G., 2011. Why are de Bruijn graphs useful for genome assembly? Nat. Biotechnol. 29, 987. https://doi.org/10.1038/NBT.2023

Cummings, D.E., Purnell, J.Q., Frayo, R.S., Schmidova, K., Wisse, B.E., Weigle, D.S., 2001. A Preprandial Rise in Plasma Ghrelin Levels Suggests a Role in Meal Initiation in Humans. Diabetes 50, 1714–1719. https://doi.org/10.2337/diabetes.50.8.1714

Dick, G.J., Andersson, A.F., Baker, B.J., Simmons, S.L., Thomas, B.C., Yelton, A.P., Banfield, J.F., 2009. Community-wide analysis of microbial genome sequence signatures. Genome Biol. 2009 108 10, 1–16. https://doi.org/10.1186/GB-2009-10-8-R85

Ebrahim, A., Lerman, J.A., Palsson, B.O., Hyduke, D.R., 2013. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. BMC Syst. Biol. 7, 74. https://doi.org/10.1186/1752-0509-7-74

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C., 2003. Isometric Logratio Transformations for Compositional Data Analysis. Math. Geol. 35, 279–300. https://doi.org/10.1023/A:1023818214614

El Homsi, M., Ducroc, R., Claustre, J., Jourdan, G., Gertler, A., Estienne, M., Bado, A., Scoazec, J.Y., Plaisancié, P., 2007. Leptin modulates the expression of secreted and membrane-associated mucins in colonic epithelial cells by targeting PKC, PI3K, and MAPK pathways. Am. J. Physiol. - Gastrointest. Liver Physiol. 293. https://doi.org/10.1152/ajpgi.00091.2007

Emms, D.M., Kelly, S., 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019 201 20, 1–14. https://doi.org/10.1186/S13059-019-1832-Y

Erb, I., Notredame, C., 2016. How should we measure proportionality on relative gene expression data? Theory Biosci. 135, 21–36. https://doi.org/10.1007/s12064-015-0220-8

Espinoza, J.L., Dupont, C.L., O'Rourke, A., Beyhan, S., Morales, P., Spoering, A., Meyer, K.J., Chan, A.P., Choi, Y., Nierman, W.C., Lewis, K., Nelson, K.E., 2021. Predicting antimicrobial mechanism-of-action from transcriptomes: A generalizable explainable artificial intelligence approach. PLOS Comput. Biol. 17, e1008857. https://doi.org/10.1371/journal.pcbi.1008857

Espinoza, J.L., Harkins, D.M., Torralba, M., Gomez, A., Highlander, S.K., Jones, M.B., Leong, P., Saffery, R., Bockmann, M., Kuelbs, C., Inman, J.M., Hughes, T., Craig, J.M., Nelson, K.E., Dupont, C.L., 2018. Supragingival Plaque Microbiome Ecology and Functional Potential in the Context of Health and Disease. MBio 9. https://doi.org/10.1128/mBio.01631-18

Espinoza, J.L., Shah, N., Singh, S., Nelson, K.E., Dupont, C.L., 2020. Applications of weighted association networks applied to compositional data in biology. Environ.

Microbiol. 22, 3020–3038. https://doi.org/10.1111/1462-2920.15091

Fernandes, A.D., Macklaim, J.M., Linn, T.G., Reid, G., Gloor, G.B., 2013. ANOVA-Like Differential Expression (ALDEx) Analysis for Mixed Population RNA-Seq. PLoS One 8, e67019. https://doi.org/10.1371/journal.pone.0067019

Fernandes, A.D., Reid, J.N.S., Macklaim, J.M., McMurrough, T.A., Edgell, D.R., Gloor, G.B., 2014. Unifying the analysis of high-throughput sequencing datasets: Characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome 2, 15. https://doi.org/10.1186/2049-2618-2-15

Finucane, M.M., Sharpton, T.J., Laurent, T.J., Pollard, K.S., 2014. A Taxonomic Signature of Obesity in the Microbiome? Getting to the Guts of the Matter. PLoS One 9, e84689. https://doi.org/10.1371/journal.pone.0084689

Franzosa, E.A., McIver, L.J., Rahnavard, G., Thompson, L.R., Schirmer, M., Weingart, G., Lipson, K.S., Knight, R., Caporaso, J.G., Segata, N., Huttenhower, C., 2018. Species-level functional profiling of metagenomes and metatranscriptomes. Nat. Methods 15, 962–968. https://doi.org/10.1038/s41592-018-0176-y

Friedman, J., Alm, E.J., 2012. Inferring Correlation Networks from Genomic Survey Data. PLoS Comput. Biol. 8, e1002687. https://doi.org/10.1371/journal.pcbi.1002687

Fuller, T.F., Ghazalpour, A., Aten, J.E., Drake, T.A., Lusis, A.J., Horvath, S., 2007. Weighted gene coexpression network analysis strategies applied to mouse weight. Mamm. Genome 18, 463–472. https://doi.org/10.1007/s00335-007-9043-3

Gao, J., Li, D., Havlin, S.,. From a single network to a network of networks. Natl. Sci.

Rev. 1.

Garza, D.R., Dutilh, B.E., 2015. From cultured to uncultured genome sequences: metagenomics and modeling microbial ecosystems. Cell. Mol. Life Sci. 72, 4287. https://doi.org/10.1007/S00018-015-2004-1

Gilbert, J.A., Blaser, M.J., Caporaso, J.G., Jansson, J.K., Lynch, S.,Knight, R. 2018. Current understanding of the human microbiome. Nature Medicine. 24:4 24(4), 392–400. https://doi.org/10.1038/nm.4517

Gloor, G.B., Macklaim, J.M., Pawlowsky-Glahn, V., Egozcue, J.J., 2017. Microbiome Datasets Are Compositional: And This Is Not Optional. Front. Microbiol. 8, 2224. https://doi.org/10.3389/fmicb.2017.02224

Gomez, A., Espinoza, J.L., Harkins, D.M., Leong, P., Saffery, R., Bockmann, M., Torralba, M., Kuelbs, C., Kodukula, R., Inman, J., Hughes, T., Craig, J.M., Highlander, S.K., Jones, M.B., Dupont, C.L., Nelson, K.E., 2017. Host Genetic Control of the Oral Microbiome in Health and Disease. Cell Host Microbe 22, 269-278.e3. https://doi.org/10.1016/j.chom.2017.08.013

Gough, E.K., Stephens, D.A., Moodie, E.E.M., Prendergast, A.J., Stoltzfus, R.J., Humphrey, J.H., Manges, A.R., 2015. Linear growth faltering in infants is associated with Acidaminococcus sp. and community-level changes in the gut microbiota. Microbiome 3. https://doi.org/10.1186/s40168-015-0089-2

Grases-Pintó, B., Abril-Gil, M., Castell, M., Rodríguez-Lagunas, M.J., Burleigh, S., Fåk Hållenius, F., Prykhodko, O., Pérez-Cano, F.J., Franch, À., 2019. Influence of Leptin and Adiponectin Supplementation on Intraepithelial Lymphocyte and Microbiota Composition in Suckling Rats. Front. Immunol. 10, 2369.

https://doi.org/10.3389/fimmu.2019.02369

Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlimi, A., Roux, S., Darzi, Y., Audic, S., Berline, L., Brum, J.R., Coelho, L.P., Espinoza, J.C.I., Malviya, S., Sunagawa, S., Dimier, C., Kandels-Lewis, S., Picheral, M., Poulain, J., Searson, S., Stemmann, L., Not, F., Hingamp, P., Speich, S., Follows, M., Karp-Boss, L., Boss, E., Ogata, H., Pesant, S., Weissenbach, J., Wincker, P., Acinas, S.G., Bork, P., De Vargas, C., Iudicone, D., Sullivan, M.B., Raes, J., Karsenti, E., Bowler, C., Gorsky, G., 2016. Plankton networks driving carbon export in the oligotrophic ocean. Nature 532, 465–470. https://doi.org/10.1038/nature16942

Gunning, D., 2017. Explainable Artificial Intelligence (XAI).

Guo, J., Bolduc, B., Zayed, A.A., Varsani, A., Dominguez-Huerta, G., Delmont, T.O., Pratama, A.A., Gazitúa, M.C., Vik, D., Sullivan, M.B., Roux, S., 2021. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. Microbiome 2021 91 9, 1–13. https://doi.org/10.1186/S40168-020-00990-Y

Hagberg, A.A., Schult, D.A., Swart, P.J., 2008. Exploring Network Structure, Dynamics, and Function using NetworkX.

Hardin, J., Mitani, A., Hicks, L., VanKoten, B., 2007. A robust measure of correlation between two genes on a microarray. BMC Bioinformatics 8, 220. https://doi.org/10.1186/1471-2105-8-220

Heimann, J.D., 2002. The extracytoplasmic function (ECF) sigma factors. Adv. Microb. Physiol. 46, 47–110. https://doi.org/10.1016/S0065-2911(02)46002-X

Hossain, M.I., Chisti, J., Yoshimatsu, S., Yasmin, R., Ahmed, T., 2015. Features in Septic Children With or Without Severe Acute Malnutrition and the Risk Factors of

Mortality. Pediatrics 135, S10–S10. https://doi.org/10.1542/peds.2014-3330q

Hsu, C.-L., Juan, H.-F., Huang, H.-C., 2015. Functional Analysis and Characterization of Differential Coexpression Networks. Sci. Rep. 5, 13295. https://doi.org/10.1038/srep13295

Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., Butterfield, C.N., Hernsdorf, A.W., Amano, Y., Ise, K., Suzuki, Y., Dudek, N., Relman, D.A., Finstad, K.M., Amundson, R., Thomas, B.C., Banfield, J.F., 2016. A new view of the tree of life. Nat. Microbiol. 1, 16048. https://doi.org/10.1038/nmicrobiol.2016.48

Hunninghake, G.W., Doerschug, K.C., Nymon, A.B., Schmidt, G.A., Meyerholz, D.K., Ashare, A., 2010. Insulin-like growth factor-1 levels contribute to the development of bacterial translocation in sepsis. Am. J. Respir. Crit. Care Med. 182, 518–525. https://doi.org/10.1164/rccm.200911-1757OC

Hutchison, C.A., Chuang, R.-Y., Noskov, V.N., Assad-Garcia, N., Deerinck, T.J., Ellisman, M.H., Gill, J., Kannan, K., Karas, B.J., Ma, L., Pelletier, J.F., Qi, Z.-Q., Richter, R.A., Strychalski, E.A., Sun, L., Suzuki, Y., Tsvetanova, B., Wise, K.S., Smith, H.O., Glass, J.I., Merryman, C., Gibson, D.G., Venter, J.C., 2016. Design and synthesis of a minimal bacterial genome. Science (80-.). 351, aad6253–aad6253. https://doi.org/10.1126/science.aad6253

Hutter, B., Schaab, C., Albrecht, S., Borgmann, M., Brunner, N.A., Freiberg, C., Ziegelbauer, K., Rock, C.O., Ivanov, I., Loferer, H., 2004. Prediction of mechanisms of action of antibacterial compounds by gene expression profiling. Antimicrob. Agents Chemother. 48, 2838–2844. https://doi.org/10.1128/AAC.48.8.2838-

2844.2004

Imai, Y., Meyer, K.J., Iinishi, A., Favre-Godal, Q., Green, R., Manuse, S., Caboni, M., Mori, M., Niles, S., Ghiglieri, M., Honrao, C., Ma, X., Guo, J., Makriyannis, A., Linares-Otoya, L., Böhringer, N., Wuisan, Z.G., Kaur, H., Wu, R., Mateus, A., Typas, A., Savitski, M.M., Espinoza, J.L., O'Rourke, A., Nelson, K.E., Hiller, S., Noinaj, N., Schäberle, T.F., D'Onofrio, A., Lewis, K., 2019. A new antibiotic selectively kills Gram-negative pathogens. Nature. https://doi.org/10.1038/s41586-019-1791-1

Jacomy, M., Venturini, T., Heymann, S., Bastian, M., 2014. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. PLoS One 9, e98679. https://doi.org/10.1371/journal.pone.0098679

Jain, C., Rodriguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T., Aluru, S., 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat. Commun. 2018 91 9, 1–8. https://doi.org/10.1038/s41467-018-07641-9

Jones, K.D.J., Berkley, J.A., 2014. Severe acute malnutrition and infection. Paediatr. Int. Child Health 34, S1–S29. https://doi.org/10.1179/2046904714Z.000000000218

Kanasi, E., Dewhirst, F.E., Chalmers, N.I., Kent, R., Jr., Moore, A., Hughes, C.V., Pradhan, N., Loo, C.Y., Tanner, A.C.R., 2010. Clonal Analysis of the Microbiota of Severe Early Childhood Caries. Caries Res. 44, 485. https://doi.org/10.1159/000320158

Kang, D.D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., Wang, Z., 2019. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction

from metagenome assemblies. PeerJ 7. https://doi.org/10.7717/PEERJ.7359

Keijser, B.J.F., E. Zaura, S.M. Huse, J.M.B.M. van der Vossen, F.H.J. Schuren, R.C. Montijn, J.M. ten Cate, W. Crielaard., 2008. Pyrosequencing analysis of the oral microflora of healthy adults. J. Dent. Res. 87, 1016–1020. https://doi.org/10.1177/154405910808701104

Kieft, K., Zhou, Z., Anantharaman, K., 2020. VIBRANT: Automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. Microbiome 8. https://doi.org/10.1186/S40168-020-00867-0

King, A.C., Wu, L., 2009. Macromolecular Synthesis and Membrane Perturbation Assays for Mechanisms of Action Studies of Antimicrobial Agents. Curr. Protoc. Pharmacol. 47, 13A.7.1-13A.7.23. https://doi.org/10.1002/0471141755.PH13A07S47@10.1002/(ISSN)1934-8290.ANTIINFECTIVES

Knights, D., Costello, E.K., Knight, R., 2011. Supervised classification of human microbiota. FEMS Microbiol. Rev. 35, 343–359. https://doi.org/10.1111/j.1574-6976.2010.00251.x

Kojima, M., Hosoda, H., Date, Y., Nakazato, M., Matsuo, H., Kangawa, K., 1999. Ghrelin is a growth-hormone-releasing acylated peptide from stomach. Nature 402, 656–660. https://doi.org/10.1038/45230

Krzywinski, M., Birol, I., Jones, S.J., Marra, M.A., 2012. Hive plots--rational approach to visualizing networks. Brief. Bioinform. 13, 627–644. https://doi.org/10.1093/bib/bbr069

Lachmann, A., Giorgi, F.M., Lopez, G., Califano, A., 2016. ARACNe-AP: gene network

reverse engineering through adaptive partitioning inference of mutual information. Bioinformatics 32, 2233–5. https://doi.org/10.1093/bioinformatics/btw216

Laczny, C.C., Sternal, T., Plugaru, V., Gawron, P., Atashpendar, A., Margossian, H., Coronado, S., der Maaten, L., Vlassis, N., Wilmes, P., 2015. VizBin - an application for reference-independent visualization and human-augmented binning of metagenomic data. Microbiome 3, 1. https://doi.org/10.1186/s40168-014-0066-1

Langfelder, P., Horvath, S., 2012. Fast R Functions for Robust Correlations and Hierarchical Clustering. J. Stat. Softw. 46.

Levering, J., Dupont, C.L., Allen, A.E., Palsson, B.O., Zengler, K., 2017. Integrated Regulatory and Metabolic Networks of the Marine Diatom Phaeodactylum tricornutum Predict the Response to Rising CO2 Levels. mSystems 2, e00142-16. https://doi.org/10.1128/mSystems.00142-16

Liljemark, W.F., Bloomquist, C.G., Uhl, L.A., Schaffer, E.M., Wolff, L.F., Pihlstrom, B.L., Bandt, C.L., 1984. Distribution of oral Haemophilus species in dental plaque from a large adult population. Infect. Immun. 46, 778.

Ling, L.L., Schneider, T., Peoples, A.J., Spoering, A.L., Engels, I., Conlon, B.P., Mueller, A., Schäberle, T.F., Hughes, D.E., Epstein, S., Jones, M., Lazarides, L., Steadman, V.A., Cohen, D.R., Felix, C.R., Fetterman, K.A., Millett, W.P., Nitti, A.G., Zullo, A.M., Chen, C., Lewis, K., 2015. A new antibiotic kills pathogens without detectable resistance. Nature 517, 455–459. https://doi.org/10.1038/nature14098

Locey, K. J., & Lennon, J. T. 2016. Scaling laws predict global microbial diversity. Proceedings of the National Academy of Sciences of the United States of America. 113(21), 5970–5975. https://doi.org/10.1073/PNAS.1521291113

Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550. https://doi.org/10.1186/S13059-014-0550-8

Lovell, D., Pawlowsky-Glahn, V., Egozcue, J.J., Marguerat, S., Bähler, J., 2015. Proportionality: A Valid Alternative to Correlation for Relative Data. PLOS Comput. Biol. 11, e1004075. https://doi.org/10.1371/journal.pcbi.1004075

Mallawaarachchi, V., Wickramarachchi, A., Lin, Y., 2020. GraphBin: Refined binning of metagenomic contigs using assembly graphs. Bioinformatics 36, 3307–3313. https://doi.org/10.1093/BIOINFORMATICS/BTAA180

Mallawaarachchi, V.G., Wickramarachchi, A.S., Lin, Y., 2021. Improving metagenomic binning results with overlapped bins using assembly graphs. Algorithms Mol. Biol. 16, 3. https://doi.org/10.1186/s13015-021-00185-6

Mandakovic, D., Rojas, C., Maldonado, J., Latorre, M., Travisany, D., Delage, E., Bihouée, A., Jean, G., Díaz, F.P., Fernández-Gómez, B., Cabrera, P., Gaete, A., Latorre, C., Gutiérrez, R.A., Maass, A., Cambiazo, V., Navarrete, S.A., Eveillard, D., González, M., 2018. Structure and co-occurrence patterns in microbial communities under acute environmental stress reveal ecological factors fostering resilience. Sci. Rep. 8, 1–12. https://doi.org/10.1038/s41598-018-23931-0

Mandal, S., Van Treuren, W., White, R.A., Eggesbø, M., Knight, R., Peddada, S.D., 2015. Analysis of composition of microbiomes: a novel method for studying microbial composition. Microb. Ecol. Heal. Dis. 26. https://doi.org/10.3402/mehd.v26.27663

McInnes, L., Healy, J., Melville, J., 2018. UMAP: Uniform Manifold Approximation and

Projection for Dimension Reduction.

Meleshko, D., Hajirasouliha, I., Korobeynikov, A., 2021. coronaSPAdes: from biosynthetic gene clusters to RNA viral assemblies. Bioinformatics. https://doi.org/10.1093/BIOINFORMATICS/BTAB597

Meleshko, D., Mohimani, H., Tracanna, V., Hajirasouliha, I., Medema, M.H., Korobeynikov, A., Pevzner, P.A., 2019. BiosyntheticSPAdes: reconstructing biosynthetic gene clusters from assembly graphs. Genome Res. 29, 1352–1362. https://doi.org/10.1101/GR.243477.118

Million, M., Diallo, A., Raoult, D., 2017. Gut microbiota and malnutrition. Microb. Pathog. https://doi.org/10.1016/j.micpath.2016.02.003

Moraleda-Muñoz, A., Marcos-Torres, F.J., Pérez, J., Muñoz-Dorado, J., 2019. Metal-responsive RNA polymerase extracytoplasmic function (ECF) sigma factors. Mol. Microbiol. 112, 385–398. https://doi.org/10.1111/MMI.14328

Morton, J.T., Marotz, C., Washburne, A., Silverman, J., Zaramela, L.S., Edlund, A., Zengler, K., Knight, R., 2019. Establishing microbial composition measurement standards with reference frames. Nat. Commun. 10, 1–11. https://doi.org/10.1038/s41467-019-10656-5

Morton, J.T., Sanders, J., Quinn, R.A., McDonald, D., Gonzalez, A., Vázquez-Baeza, Y., Navas-Molina, J.A., Song, S.J., Metcalf, J.L., Hyde, E.R., Lladser, M., Dorrestein, P.C., Knight, R., 2017. Balance Trees Reveal Microbial Niche Differentiation. mSystems 2, e00162-16. https://doi.org/10.1128/mSystems.00162-16

Müller, T.D., Nogueiras, R., Andermann, M.L., Andrews, Z.B., Anker, S.D., Argente, J., Batterham, R.L., Benoit, S.C., Bowers, C.Y., Broglio, F., Casanueva, F.F.,

D'Alessio, D., Depoortere, I., Geliebter, A., Ghigo, E., Cole, P.A., Cowley, M., Cummings, D.E., Dagher, A., Diano, S., Dickson, S.L., Diéguez, C., Granata, R., Grill, H.J., Grove, K., Habegger, K.M., Heppner, K., Heiman, M.L., Holsen, L., Holst, B., Inui, A., Jansson, J.O., Kirchner, H., Korbonits, M., Laferrère, B., LeRoux, C.W., Lopez, M., Morin, S., Nakazato, M., Nass, R., Perez-Tilve, D., Pfluger, P.T., Schwartz, T.W., Seeley, R.J., Sleeman, M., Sun, Y., Sussel, L., Tong, J., Thorner, M.O., van der Lely, A.J., van der Ploeg, L.H.T., Zigman, J.M., Kojima, M., Kangawa, K., Smith, R.G., Horvath, T., Tschöp, M.H., 2015. Ghrelin. Mol. Metab. https://doi.org/10.1016/j.molmet.2015.03.005

National Center for Biotechnology Information, (NCBI), 2020. SRA Data Format and Storage [WWW Document]. URL https://ncbiinsights.ncbi.nlm.nih.gov/tag/sra/ (accessed 10.18.21).

Nayfach, S., Camargo, A.P., Schulz, F., Eloe-Fadrosh, E., Roux, S., Kyrpides, N.C., 2020. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat. Biotechnol. 2020 395 39, 578–585. https://doi.org/10.1038/s41587-020-00774-7

Nayfach, S., Rodriguez-Mueller, B., Garud, N., Pollard, K.S., 2016. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. Genome Res. 26, 1612–1625. https://doi.org/10.1101/gr.201863.115

NHGRI, n.d. DNA Sequencing Costs: Data [WWW Document]. 2020. URL https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data (accessed 4.29.20).

Nissen, J.N., Johansen, J., Allesøe, R.L., Sønderby, C.K., Armenteros, J.J.A., Grønbech, C.H., Jensen, L.J., Nielsen, H.B., Petersen, T.N., Winther, O., Rasmussen, S., 2021. Improved metagenome binning and assembly using deep variational autoencoders. Nat. Biotechnol. 2021 395 39, 555–560. https://doi.org/10.1038/s41587-020-00777-4

Nurk, S., Meleshko, D., Korobeynikov, A., Pevzner, P.A., 2017. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 27, 824–834. https://doi.org/10.1101/gr.213959.116

O'Rourke, A., Beyhan, S., Choi, Y., Morales, P., Chan, A.P., Espinoza, J.L., Dupont, C.L., Meyer, K.J., Spoering, A., Lewis, K., Nierman, W.C., Nelson, K.E., 2020. Mechanism-of-action classification of antibiotics by global transcriptome profiling. Antimicrob. Agents Chemother. 64. https://doi.org/10.1128/AAC.01207-19

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., Tyson, G.W., 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 25, 1043–55. https://doi.org/10.1101/gr.186072.114

Pascual-García, A., Bonhoeffer, S., Bell, T., 2020. Metabolically cohesive microbial consortia and ecosystem functioning. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 375. https://doi.org/10.1098/RSTB.2019.0245

Paulson, J.N., Stine, O.C., Bravo, H.C., Pop, M., 2013. Differential abundance analysis for microbial marker-gene surveys. Nat. Methods 10, 1200–1202. https://doi.org/10.1038/nmeth.2658

Pawlowsky-Glahn, V., Egozcue, J., Tolosana-Delgado, R., 2015. Modeling and analysis

of compositional data.

Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R., 2011. Lecture Notes on Compositional Data Analysis.

Platts-Mills, J.A., Taniuchi, M., Uddin, M.J., Sobuz, S.U., Mahfuz, M., Gaffar, S.M.A., Mondal, D., Hossain, M.I., Islam, M.M., Ahmed, A.M.S., Petri, W.A., Haque, R., Houpt, E.R., Ahmed, T., 2017. Association between enteropathogens and malnutrition in children aged 6-23 mo in Bangladesh: A case-control study. Am. J. Clin. Nutr. 105, 1132–1138. https://doi.org/10.3945/ajcn.116.138800

Quinn, T. P., Crowley, T.M., Richardson, M.F., 2018. Benchmarking differential expression analysis tools for RNA-Seq: Normalization-based vs. log-ratio transformation-based methods. BMC Bioinformatics 19. https://doi.org/10.1186/s12859-018-2261-8

Quinn, T.P., Erb, I., 2020. Amalgams : data-driven amalgamation for the reference-free dimensionality reduction of zero-laden compositional data 1–16.

Quinn, T.P., Erb, I., Richardson, M.F., Crowley, T.M., 2018. Understanding sequencing data as compositions: an outlook and review. Bioinformatics 34, 2870–2878. https://doi.org/10.1093/bioinformatics/bty175

Quinn, T.P., Richardson, M.F., Lovell, D., Crowley, T.M., 2017. Propr: An R-package for Identifying Proportionally Abundant Features Using Compositional Data Analysis. Sci. Rep. 7, 1–9. https://doi.org/10.1038/s41598-017-16520-0

Razumov, A.S., 1932. The direct method of calculation of bacteria in water: comparison with the Koch method. Mikrobiologija 1, 131–146.

Ren, J., Ahlgren, N.A., Lu, Y.Y., Fuhrman, J.A., Sun, F., 2017. VirFinder: a novel k-mer

based tool for identifying viral sequences from assembled metagenomic data. Microbiome 5, 69. https://doi.org/10.1186/s40168-017-0283-5

Rivera-Pinto, J., Egozcue, J.J., Pawlowsky-Glahn, V., Paredes, R., Noguera-Julian, M., Calle, M.L., 2018. Balances: a New Perspective for Microbiome Analysis. mSystems 3. https://doi.org/10.1128/msystems.00053-18

Robinson, M.D., McCarthy, D.J., Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140. https://doi.org/10.1093/bioinformatics/btp616

Santoro, E.P., Borges, R.M., Espinoza, J.L., Freire, M., Messias, C.S.M.A., Villela, H.D.M., Pereira, L.M., Vilela, C.L.S., Rosado, J.G., Cardoso, P.M., Rosado, P.M., Assis, J.M., Duarte, G.A.S., Perna, G., Rosado, A.S., Macrae, A., Dupont, C.L., Nelson, K.E., Sweet, M.J., Voolstra, C.R., Peixoto, R.S., 2021. Coral microbiome manipulation elicits metabolic and genetic restructuring to mitigate heat stress and evade mortality. Sci. Adv. 7. https://doi.org/10.1126/sciadv.abg3088

Schüssler-Fiorenza Rose, S.M., Contrepois, K., Moneghetti, K.J., Zhou, W., Mishra, T., Mataraso, S., Dagan-Rosenfeld, O., Ganz, A.B., Dunn, J., Hornburg, D., Rego, S., Perelman, D., Ahadi, S., Sailani, M.R., Zhou, Y., Leopold, S.R., Chen, J., Ashland, M., Christle, J.W., Avina, M., Limcaoco, P., Ruiz, C., Tan, M., Butte, A.J., Weinstock, G.M., Slavich, G.M., Sodergren, E., McLaughlin, T.L., Haddad, F., Snyder, M.P., 2019. A longitudinal big data approach for precision health. Nat. Med. 25, 792–804. https://doi.org/10.1038/s41591-019-0414-6

Schwarzer, M., Makki, K., Storelli, G., Machuca-Gayet, I., Srutkova, D., Hermanova, P., Martino, M.E., Balmand, S., Hudcovic, T., Heddi, A., Rieusset, J., Kozakova, H.,

Vidal, H., Leulier, F., 2016. Lactobacillus plantarum strain maintains growth of infant mice during chronic undernutrition. Science (80-.). 351, 854–857. https://doi.org/10.1126/science.aad8588

Sender, R., Fuchs, S., Milo, R., 2016a. Revised Estimates for the Number of Human and Bacteria Cells in the Body. PLOS Biol. 14, e1002533. https://doi.org/10.1371/JOURNAL.PBIO.1002533

Sender, R., Fuchs, S., Milo, R., 2016b. Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans. Cell 164, 337–340. https://doi.org/10.1016/j.cell.2016.01.013

Shaiber, A., Eren, A.M., 2019. Composite metagenome-assembled genomes reduce the quality of public genome repositories. MBio 10. https://doi.org/10.1128/MBIO.00725-19

Shomorony, I., Cirulli, E.T., Huang, L., Napier, L.A., Heister, R.R., Hicks, M., Cohen, I. V., Yu, H.C., Swisher, C.L., Schenker-Ahmed, N.M., Li, W., Nelson, K.E., Brar, P., Kahn, A.M., Spector, T.D., Caskey, C.T., Venter, J.C., Karow, D.S., Kirkness, E.F., Shah, N., 2020. An unsupervised learning approach to identify novel signatures of health and disease from multimodal data. Genome Med. 12, 7. https://doi.org/10.1186/s13073-019-0705-z

Sieber, C.M.K., Probst, A.J., Sharrar, A., Thomas, B.C., Hess, M., Tringe, S.G., Banfield, J.F., 2018. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nat. Microbiol. 2018 37 3, 836–843. https://doi.org/10.1038/s41564-018-0171-1

Silverman, J.D., Washburne, A.D., Mukherjee, S., David, L.A., 2017. A phylogenetic

transform enhances analysis of compositional microbiota data. Elife 6. https://doi.org/10.7554/eLife.21887

Singh, B.K., Trivedi, P., Egidi, E., Macdonald, C.A., Delgado-Baquerizo, M., 2020. Crop microbiome and sustainable agriculture. Nat. Rev. Microbiol. 2020 1811 18, 601–602. https://doi.org/10.1038/s41579-020-00446-y

Slinger, R., Lau, K., Slinger, M., Moldovan, I., Chan, F., 2017. Higher atypical enteropathogenic Escherichia coli (a-EPEC) bacterial loads in children with diarrhea are associated with PCR detection of the EHEC factor for adherence 1/lymphocyte inhibitory factor A (efa1/lifa) gene. Ann. Clin. Microbiol. Antimicrob. 16, 16. https://doi.org/10.1186/s12941-017-0188-y

Smith, M.I., Yatsunenko, T., Manary, M.J., Trehan, I., Mkakosya, R., Cheng, J., Kau, A.L., Rich, S.S., Concannon, P., Mychaleckyj, J.C., Liu, J., Houpt, E., Li, J. V., Holmes, E., Nicholson, J., Knights, D., Ursell, L.K., Knight, R., Gordon, J.I., 2013. Gut microbiomes of Malawian twin pairs discordant for kwashiorkor. Science (80-.). 339, 548–554. https://doi.org/10.1126/science.1229000

Song, L., Langfelder, P., Horvath, S., 2012. Comparison of co-expression measures: mutual information, correlation, and model based indices. BMC Bioinformatics 13, 328. https://doi.org/10.1186/1471-2105-13-328

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P., 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. 102, 15545–15550. https://doi.org/10.1073/pnas.0506580102

Subramanian, S., Huq, S., Yatsunenko, T., Haque, R., Mahfuz, M., Alam, M.A., Benezra, A., Destefano, J., Meier, M.F., Muegge, B.D., Barratt, M.J., VanArendonk, L.G., Zhang, Q., Province, M.A., Petri, W.A., Ahmed, T., Gordon, J.I., 2014. Persistent gut microbiota immaturity in malnourished Bangladeshi children. Nature 510, 417–421. https://doi.org/10.1038/nature13421

Takahashi, N., Washio, J., Mayanagi, G., 2010. Metabolomics of supragingival plaque and oral bacteria. J. Dent. Res. 89, 1383–1388. https://doi.org/10.1177/0022034510377792

Luckey, T.D., 1972. Introduction to intestinal microecology. Am. J. Clin. Nutr. 25, 1292–1294. https://doi.org/10.1093/AJCN/25.12.1292

Tien, M.-T., Girardin, S.E., Regnault, B., Le Bourhis, L., Dillies, M.-A., Coppée, J.-Y., Bourdet-Sicard, R., Sansonetti, P.J., Pédron, T., 2006. Anti-Inflammatory Effect of Lactobacillus casei on Shigella -Infected Human Intestinal Epithelial Cells . J. Immunol. 176, 1228–1237. https://doi.org/10.4049/jimmunol.176.2.1228

Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C.M., Knight, R., Gordon, J.I., 2007. The Human Microbiome Project. Nat. 2007 4497164 449, 804–810. https://doi.org/10.1038/nature06244

Van Der Maaten, L., 2014. Accelerating t-SNE using Tree-Based Algorithms. J. Mach. Learn. Res. 15, 1–21.

Villaverde, A.F., Ross, J., Morán, F., Banga, J.R., 2014. MIDER: Network inference with mutual information distance and entropy reduction. PLoS One 9, e96732. https://doi.org/10.1371/journal.pone.0096732

Voorhies, A.A., Mark Ott, C., Mehta, S., Pierson, D.L., Crucian, B.E., Feiveson, A.,

Oubre, C.M., Torralba, M., Moncera, K., Zhang, Y., Zurek, E., Lorenzi, H.A., 2019. Study of the impact of long-duration space missions at the International Space Station on the astronaut microbiome. Sci. Rep. 9. https://doi.org/10.1038/s41598-019-46303-8

Washburne, A.D., Silverman, J.D., Leff, J.W., Bennett, D.J., Darcy, J.L., Mukherjee, S., Fierer, N., David, L.A., 2017. Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. PeerJ 5, e2969. https://doi.org/10.7717/peerj.2969

Weiss, H., Hertzberg, V.S., Dupont, C., Espinoza, J.L., Levy, S., Nelson, K., Norris, S., 2018. The Airplane Cabin Microbiome. Microb. Ecol. https://doi.org/10.1007/s00248-018-1191-3

Wood, D.E., Lu, J., Langmead, B., 2019. Improved metagenomic analysis with Kraken 2. Genome Biol. 20, 1–13. https://doi.org/10.1186/S13059-019-1891-0/FIGURES/2

Wu, Y.-W., Simmons, B.A., Singer, S.W., 2016. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics 32, 605–607. https://doi.org/10.1093/BIOINFORMATICS/BTV638

Yang, X., He, L., Yan, S., Chen, X., Que, G., 2021. The impact of caries status on supragingival plaque and salivary microbiome in children with mixed dentition: a cross-sectional survey. BMC Oral Heal. 2021 211 21, 1–13. https://doi.org/10.1186/S12903-021-01683-0

Yu, G., Jiang, Y., Wang, J., Zhang, H., Luo, H., 2018. BMC3C: binning metagenomic contigs using codon usage, sequence composition and read coverage. Bioinformatics 34, 4172–4179. https://doi.org/10.1093/BIOINFORMATICS/BTY519

Zampieri, M., Szappanos, B., Buchieri, M.V., Trauner, A., Piazza, I., Picotti, P., Gagneux, S., Borrell, S., Gicquel, B., Lelievre, J., Papp, B., Sauer, U., 2018. High-throughput metabolomic analysis predicts mode of action of uncharacterized antimicrobial compounds. Sci. Transl. Med. 10. https://doi.org/10.1126/scitranslmed.aal3973

Zaura, E., Keijser, B.J., Huse, S.M., Crielaard, W., 2009. Defining the healthy "core microbiome" of oral microbial communities. BMC Microbiol. 9, 259. https://doi.org/10.1186/1471-2180-9-259

Zhang, L., Andersen, D., Roager, H.M., Bahl, M.I., Hansen, C.H.F., Danneskiold-Samsøe, N.B., Kristiansen, K., Radulescu, I.D., Sina, C., Frandsen, H.L., Hansen, A.K., Brix, S., Hellgren, L.I., Licht, T.R., 2017. Effects of Gliadin consumption on the Intestinal Microbiota and Metabolic Homeostasis in Mice Fed a High-fat Diet. Sci. Rep. 7, 44613. https://doi.org/10.1038/srep44613

Zhou, W., Sailani, M.R., Contrepois, K., Zhou, Y., Ahadi, S., Leopold, S.R., Zhang, M.J., Rao, V., Avina, M., Mishra, T., Johnson, J., Lee-McMullen, B., Chen, S., Metwally, A.A., Tran, T.D.B., Nguyen, H., Zhou, X., Albright, B., Hong, B.-Y., Petersen, L., Bautista, E., Hanson, B., Chen, L., Spakowicz, D., Bahmani, A., Salins, D., Leopold, B., Ashland, M., Dagan-Rosenfeld, O., Rego, S., Limcaoco, P., Colbert, E., Allister, C., Perelman, D., Craig, C., Wei, E., Chaib, H., Hornburg, D., Dunn, J., Liang, L., Rose, S.M.S.-F., Kukurba, K., Piening, B., Rost, H., Tse, D., McLaughlin, T., Sodergren, E., Weinstock, G.M., Snyder, M., 2019. Longitudinal multi-omics of host–microbe dynamics in prediabetes. Nature 569, 663–671. https://doi.org/10.1038/s41586-019-1236-x

Zoffmann, S., Vercruysse, M., Benmansour, F., Maunz, A., Wolf, L., Blum Marti, R., Heckel, T., Ding, H., Truong, H.H., Prummer, M., Schmucki, R., Mason, C.S., Bradley, K., Jacob, A.I., Lerner, C., Araujo del Rosario, A., Burcin, M., Amrein, K.E., Prunotto, M., 2019. Machine learning-powered antibiotics phenotypic drug discovery. Sci. Rep. 9, 1–14. https://doi.org/10.1038/s41598-019-39387-9