Metagenomic systems biology: frameworks for modeling and characterizing the gut microbiome

Sharon I. Greenblum

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2014

Reading Committee:

Elhanan Borenstein, Chair

David Fredericks

Joshua M. Akey

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

Metagenomic systems biology: Systems-level modeling and characterization of the gut microbiome

Sharon I. Greenblum

Chair of the Supervisory Committee:
Elhanan Borenstein, Associate Professor
Department of Genome Sciences

Though invisible to the naked eye, microbes are crucial to life as we know it. These tiny single-celled organisms are found in almost every known environment, helping to maintain balance across a vast array of ecological niches plays. Within each site, microbes may form intricate multi-species communities capable of carrying out diverse and complex metabolic processes. The set of microbes inhabiting the human gut (the human gut microbiome) comprises one of the richest and most well-studied of these communities, and shifts in the composition of this microbiome have been shown to have significant implications for host health. However, while current comparative studies mostly focus on characterizing gut microbiomes in terms of the relative abundance of individual species or genes, such profiles

offer limited translation to overall community capabilities, and may thus offer limited predictive capacity for effect on the host. Here, I develop frameworks for characterizing and comparing microbiomes as integrated systems, leveraging concepts from systems biology to provide a deeper context for interpreting differences in community composition. In chapter 1, I describe current efforts to characterize microbial communities and the potential advantages of a systems-level perspective. In chapter 2, I present a method for constructing and characterizing topological network models of microbial community metabolism, and then identify specific topological differences between human gut communities from healthy, obese, and IBD-afflicted individuals. The results suggest that the gut environment plays a critical role in shaping microbiome topology, or structure. In chapter 3, I examine gut communities from host species across the mammalian phylogenetic tree and identify groups of functionally-related genes that co-occur across hosts. I term these gene groups 'assembly modules', and demonstrate their value for understanding the functional units of microbiome assembly and adaptation. In chapter 4, I relate differences in community function back to individual microbial strains, focusing on functions whose representation across organisms within a given species is community-dependent. Establishing a computational pipeline to detect these strain-specific functions, and generating a database of their frequency across 109 human gut microbiomes, I show that strain-specific functions are widespread among species associated with the gut environment, and that some of the most prominent, such as virulence, antibiotic resistance, and nutrient transport, may have significance for host-microbiome stability. Finally, in chapter 5, I offer some perspective on how the systems-level frameworks presented here may be used in future studies of microbial communities, potentially incorporating burgeoning new technologies and growing data resources, and how continued work in this vein may advance our understanding of the microbial world in relation to our own.

Acknowledgements

I would like thank my family and friends for keeping me balanced and giving me ample opportunities to not only study life, but enjoy it. Special thanks to my Eastlake sisters and the Sock Mafia family for being there, always, and for all of the adventures along the way. Extra special thanks to my mom for listening to me reflect and fret from afar and for being an unwavering source of trust, support, and excellent advice.

I would like to thank my labmates and classmates for being all-around amazing and encouraging people. Their positivity and comradery were an unexpectedly wonderful part of this journey.

Finally, I would like to thank my advisor Elhanan. As a scientist, his vision and creativity punctuated every conversation, and his has perspective shaped the way I think about the world. As a mentor his door was always open, and his honest opinion, undivided attention, and a neatly bulleted list of constructive feedback were always close at hand. And as a role model, his work ethic and integrity have been truly inspiring. This work would not have been possible without his guidance and patience.

Table of Contents

List of Figures and Tables

1. Introduction

1.1 Know your neighbor: Challenges in the study of microbial communities

Look around. A quick glimpse at our surroundings reveals a world in motion – constant change, growth, and interaction. It is this macroscopic view that has so delighted and intrigued scientists, artists, and children alike. But beneath this world lies another, invisible to the human eye, yet teeming with life. Tiny one-celled organisms – microbes – cover literally every surface imaginable. They are our most distant ancestors, primarily members of the domains Bacteria and Archaea, which diverged from our own lineage more than 2 billion years ago. Today though, we live with them in relative harmony, sharing space and resources in the pursuit of a largely mutually beneficial co-existence. While we move through our daily lives, they are hard at work as well, maintaining ecological equilibrium everywhere from the depths of the oceans to the deepest reaches of our stomachs. Accordingly however, disruption in the microbial world can have drastic and rippling consequences in our own. Understanding the precise ways in which these worlds are intertwined has been an ongoing challenge for centuries, yet it is only within the past few decades that technical advances have allowed us to study microbes in detail, in their natural habitats. In the following section I outline what has been learned from recent studies, concentrating specifically on the microbes inhabiting the human gut as an environment with particular clinical relevance. I briefly describe the methodologies currently used to characterize and compare human gut microbial communities, as well as their limitations.

1.1.1 The importance of human gut microbial communities

Microbes may be the world's best colonizers. With their territory extending to the far corners of the globe, microbes have devised ways to profit from the scantiest bit of

resources in the most inhospitable of locales[1]. Their ubiquity can in large part be attributed to their incredible diversity. Recent estimates have placed the number of microbial species on earth in the range of 10^7-10^9 [2], collectively encompassing a vast array of metabolic capabilities and reflecting adaptation to a wide range of environmental niches.

In many cases, microbes do not merely exist, but are actually vital to the health and stability of their environment. Co-evolution has yielded many examples in which microbes and their environment become locked in a delicate balance, with microbes reliant on external resources provided by the environment in exchange for microbial metabolic capabilities lacking in the host. For example, plant-associated microbes such as *Rhizobia spp*. rely on plant-derived carbohydrates for energy and in turn provide their hosts with sources of fixed nitrogen and phosphate [3], while microbes in the cow rumen break down the cellulose in consumed grasses, making energy and nutrients available to the host [4]. The absence of active and appropriate microbes can be acutely harmful. For example, experiments with germ-free or gnotobiotic mice have shown that these mice have reduced metabolic activity and are more prone to immunologic disorders than their colonized counterparts [5], [6], while gnotobiotic plants suffer from reduced viability, reproductive capacity, and resistance to contamination [7].

A microbial niche long considered one of the most critically important for human health is the human gut. A single human gut may harbor over a hundred trillion microbial cells, outnumbering human cells ten-to-one [8]. This collection of microbes, collectively referred to as the human gut microbiome[9], may comprise thousands of co-existing species [10]. Notably, many of these microbial organisms do not merely co-exist, but have been shown to engage in complex interactions[11], [12]. Essentially, the set of resident species form a community, relying on each other for a variety of metabolic, signaling, and ,transport-

related functions, and facilitating complex community-wide processes that would be unachievable by a single species alone [13], [14].

Over human evolution, we have come to rely on this diverse set of microbes and the important capabilities afforded them by their complex community-based lifestyle for a number of key processes, including vitamin and amino acid biosynthesis, dietary energy harvest, and immune development [15], [16]. However, the impact of a microbiome on its host is modulated by its composition, as the exact set of species present and their relative amounts can have significant implications for what the community as a whole is capable of doing. For example, changing community composition via transplant of an intact gut microbiome from one individual to another has repeatedly been shown to prompt rapid changes in human and mouse hosts, and even prompt recovery from debilitating disease [17]. It is believed that each individual has a unique compositional 'signature', determined initially upon birth when a newborn is first colonized, primarily by microbes from its mother[18], and persisting over time. However, subtle or even large-scale changes in the composition of the human gut microbiome can occur on a daily basis, dependent on the interplay of a number of complex factors such as diet, location, host exposure and activity[19]. Most strikingly though, a growing collection of studies suggest that large-scale shifts in microbiome composition may be an important and informative biomarker for the health of the human host[20]–[23].

1.1.2 Common methods for characterizing microbial community composition

Determining the composition of a given gut microbiome however, has proven challenging, long hampered by the fact that the vast majority of microbes are resistant to isolation and culturing, and thus cannot be studied using traditional methods [24]. Fortunately, the past few decades have seen the rise of next-generation sequencing, and

recent work has made use of these new technologies to advance culture-free methods of studying microbial communities. These methods focus on the sequencing of short segments of microbial DNA sampled directly from the environment, without isolation of individual species. In using such sequences to study microbiome composition, two primary modes of have become widespread: the use of marker genes to assess the diversity and relative abundance of species in the community, and shotgun metagenomics to provide a snapshot of the set of genomic functions available to the community.

1.1.1.1 Characterizing species composition via marker genes

The phylogenetic diversity of a community can be assessed by sequencing of a species-specific marker, such as the bacterial 16s ribosomal RNA (rRNA) gene. The gene contains highly conserved regions allowing primer construction and targeted sequencing, as well as species-specific hypervariable regions, distinguishing 16s sequences from different taxonomic groups [25]. Since the 16s sequences of many bacterial species are still unknown, sequences with high sequence similarity (95-99%) are often clustered into species-level groups known as operational taxonomic units (OTUs). The relative abundance of sequences in each OTU is used as a proxy for species composition, and representative sequences from some OTUs can be assigned to known phylogenies. More recently, other sets of universal single-copy marker genes have also been suggested as more robust indicators of species composition[26], though 16s analysis remains prevalent.

1.1.2.2 Characterizing functional composition via metagenomics

Species composition, however, may offer only limited insight into metabolic activity, since the genomes of a vast majority of species remain unsequenced and uncharacterized. A gene-centric, metagenomic approach uses shotgun sequencing to obtain an unbiased sample of genetic material from across an entire community. Importantly, this type of analysis is conceptually different than most work at the species

level. A microbiome is essentially viewed as a microbial 'super-organism' (17), in which species boundaries are ignored and the entire set of genes recovered from an environment is studied as if part of a single 'metagenome. The short sequences obtained are frequently on the order of 30-200 bp each depending on the sequencing technology, and sequencing is sometimes followed by assembly to obtain full genes. More sophisticated sequencing technologies can produce hundreds of gigabytes of data in a high-throughput manner, and larger scale studies have surveyed hundreds of metagenomic samples at impressive sequencing depths [27]. Genes found in the metagenome can be directly mapped to pathways and functional annotations via a growing number of public databases such as KEGG[28], SEED[29], and BioCyc[30], providing a means of capturing the metabolic potential of the community as a whole.

1.1.3 Insights from comparative studies of human gut communities

Results from marker gene and metagenomic studies have shown that species or gene composition can vary widely across microbiomes. In some cases, certain species or genes are found to differ significantly between multiple sets of microbiome samples from different environments or host states[31], [32], while other associations have been established by measuring the compositional changes following external perturbation of a controlled environment[33], [34]. A number of statistical measures have been proposed to summarize and compare microbiome composition, though agreement on a specific gold-standard has not yet been reached[35]–[37].

As awareness of the importance of the gut microbiome for human health mounts, more effort has poured into detecting specific associations between microbiome composition and the state of the human host. A number of metrics have been used to identify consistent shifts in the balance of specific species or higher order taxonomic groups that correlate with host nationality, age, diet, and disease. Studies comparing healthy and obese individuals

have been among the most prominent, a number of which have reported differences in the relative ratio of *Bacteriodetes* to *Firmicutes*[38]–[41], although in some cases, findings from similar studies are directly conflicting[42].

In parallel, metagenomic data has been mined to identify specific gene compositions that may discriminate host groups, and in some instances key differences have been identified[40]. However, whereas variation in species composition across microbiomes is often readily apparent, a number of metagenome studies have revealed high uniformity in functional repertoire[43], even in the presence of disease or altered host phenotype. Moreover, though identifying differential gene sets may provide preliminary insights into individual functions that differ between microbiomes, they often fall short of providing a comprehensive understanding of how these differences affect the capabilities of the microbial community as a whole, and consequently how such differences translate to meaningful implications for the health of the host.

1.2 Bacterial communities as biological systems

Most studies of microbial community composition thus far hinge on characterizing and comparing microbiomes as lists of parts – species or genes and their relative abundances. Yet, as empirical and experimental evidence suggests, a microbial community is truly more than the sum of its parts, metabolizing and producing complex biomolecules that are unattainable by any one species in isolation. It is the confluence of the diverse capabilities of various species that allows a microbiome to remain a stable, functional system capable of, for example, providing critical support to an entity as complex and highly evolved as the human body. Characterizing microbial systems in a framework that accounts for and highlights these systems-level features will be a crucial step towards understanding how the composition of a microbial community translates to functional consequences for the host environment.

1.2.1 The promise of systems biology

Systems biology, as a conceptual framework, entails describing any system in terms of both its parts as well as the interactions between those parts[44]. From ecological food webs to protein signaling pathways, systems are often modeled as networks of nodes and edges, with nodes representing various elements in the system and edges represent the interactions between these elements. Applying this type of framework to the system of a microbial community could include modeling the interactions between organisms, species, and/or genes across the entire community [45], [46], allowing access to the rich trove of tools that has already been developed for systems analysis. For example, recent work has shown that modeling interactions between species can provide novel insight into the forces that shape phylogenetic community composition[47].

Gene-based metabolic models, however, have primarily been applied to the study of individual microbial species. Such models have produced quantitatively-tractable genome-scale metabolic models that have been used to predict the growth rate of an organism under given culture conditions via constraint-based algorithms [48]. Topological analyses have also proven powerful for identifying features associated with various metabolic functional and evolutionary properties, including transcriptional regulation [49], genetic and environmental robustness [50], and adaptation[51]. In contrast to constraint-based analyses, which usually require detailed manually curated data, topological models (representing only the connectivity of the metabolic network) permit model construction on a very large scale.

1.2.2 Previous efforts to model bacterial communities

To date however, gene-based metabolic modeling techniques have been applied only sparingly to studies of *community* metabolism.

Instead, functional inferences about a microbial community are often obtained by mapping detected genes to known metabolic pathways. Using a variety of statistical approaches, the prevalence of certain metabolic pathways in a metagenome have been shown to correlate with various host states and environmental factors [16], [52], [53]. A number of tools have been developed to analyze and visualize metagenomic data in terms of pathways [54], [55], some utilizing pathway information to infer the presence of missing enzymes or to correct the abundance of others[56].

While such pathway-based analyses represent powerful and efficient ways to describe the functional aspects of microbiome composition, relatively few studies go beyond characterization of pathway abundances and directly take into account the relationships between the various pathways or the overall organization of the metabolic network. Combining the level of sophistication achieved in single-organism metabolic modeling techniques with the comprehensiveness of pathway-based tools is the challenge of community-level models. To achieve such models will require high-throughput analytic pipelines, accurate functional annotations, and a deeper understanding of the relationship between species and gene composition. If achieved however, community-wide models may provide valuable insight into the contribution of diverse functional elements to community metabolic potential and open the door to predictive models of the role of the microbiome within the context of an ecosystem.

1.3 Objectives

While work towards characterizing community composition has progressed at a rapid rate thanks to advancing technologies, we still lack an appropriate platform for understanding how different sets of microbial genes or species form a functional unit capable of enacting metabolic processes. Systems biology provides a conceptual framework for studying not only the constituent parts of a system, but also their interactions.

Applying systems biology concepts and methodologies to the glut of metagenomic data now available may help to obtain a more complete depiction of the forces that guide microbiome assembly and the forces that cause compositional shifts, as well as a way to assess the implications of these shifts for the state of the host. The work proposed here represents a first step in this process. In the following chapters, I present a number of investigations, each using high-throughput metagenomic data and well-established systems biology tools to explore different aspects of microbial community function. My research objectives were the following:

- *To define a modeling framework for translating community composition to microbiome functional potential.* How can we integrate both the components of a microbial community and their interactions into a straightforward yet biologically relevant model?
- *To establish meaningful metrics for characterizing and comparing the functional potential of various microbiomes.* Which aspects of microbial community models can be directly quantified, which vary most, which carry significance for the state of the host, and which have clear functional interpretations?
- *To assess the units of microbiome functional variation and evolution.* Does variation between microbiomes simply entail differences in the relative abundance of individual genes? Or can we detect coherent functionally-related gene modules that tend to co-vary across microbiomes? Does the combination of gene modules present in a microbiome give us greater insight into its functional capacity?
- *How is functional variation reflected across the genomes of individual organisms within each microbial species?* How appropriate and comprehensive are current reference genomes for predicting the functional capacities of a microbiome? Which types of functions are most predictable, and which are most dependent on community context?

2. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease.

This chapter is based on the following manuscript published in *Proceedings of the National Academy of Sciences*:

Greenblum S, Turnbaugh PJ, Borenstein E. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. Proc Natl Acad Sci USA 2012;109:594–599.

2.1 Summary

The human microbiome plays a key role in a wide range of host-related processes and has a profound effect on human health. Comparative analyses of the human microbiome have revealed substantial variation in species and gene composition associated with a variety of disease states, but fall short of providing a comprehensive understanding of the impact of this variation on the community and the host. Here, I introduce a *metagenomic systems biology* computational framework, integrating metagenomic data with a systems-level analysis and network-based in-silico models. Focusing on the gut human microbiome, I analyze fecal metagenomic data from 124 unrelated individuals as well as six monozygotic twin-pairs and their mothers, and reconstruct community-level metabolic models. Placing variations in gene

abundance in the context of these models, I identify both gene-level and network-level topological differences associated with obesity and IBD. I show that genes associated with either of these host states tend to be located at the periphery of the metabolic network, and that obesity associated genes tend to be sparsely connected and enriched for topologically derived metabolic 'inputs'. These findings suggest that microbiomes differ primarily in their interface with the host, directly affecting its metabolism. I further demonstrate that obese microbiomes are less modular, a hallmark of adaptation to a low diversity environment. I additionally link these topological variations to community species composition. The system-level approach presented here lays the foundation for a novel framework for studying the human microbiome, its organization, function, and impact on human health.

2.2 Background

We humans are mostly microbes. Microbial communities populate numerous sites in the human anatomy, together comprising over a hundred trillion microbial cells[8]. Amongst the various body habitats, the most densely colonized is the distal gut [10]. The normal gut flora alone consists of hundreds of bacterial species, collectively encoding an enormous gene set that is 150-fold larger than the set of human genes[57]. The gut microbiome plays a key role in many essential processes, including vitamin and amino acid biosynthesis, dietary energy harvest, and immune development[43], [58]. Transferring a donor microbiota into a recipient can induce various donor phenotypes (including increased capacity for energy harvest[40], increased adiposity and metabolic syndrome[59] or prompt the recovery of a sick recipient[17], suggesting a promising avenue for clinical application via directed manipulation of the microbiome. Characterizing the capacity of the human microbiome, its interaction with the host, and its contribution to various disease states has therefore the potential to provide deep insight

into both normal human physiology and human disease, and calls for a predictive systems-level understanding of community function and structure.

Addressing this challenge, world-wide research initiatives [16], [60] have recently started to map the human microbiome, providing insight into previously uncharted species and genes. Such surveys have revealed, for example, marked associations between the species composition of the gut microbiome and a variety of host phenotypes [27], [43], [61]. Species profiles, however, cannot be easily translated into function, since it is not clear how variation in the composition of species in the microbiome affects the metabolic activity of the community and consequently the host. In contrast, metagenomic shotgun sequencing of community DNA and a gene-centric comparative approach [32], [57], [62] may capture functional differences in the metabolic potential of the community. Yet, comparative metagenomic analysis of the gut microbiome has revealed only small sets of enriched genes, which provide preliminary insights into relevant functional differences, but may not provide a comprehensive, systems-level understanding of the variation and its potential effect on the host-microbiome supra-organism [63], [64].

Here, we introduce a novel framework for studying the human microbiome, integrating metagenomic data with a systems-level network analysis. This metagenomic systems biology approach goes beyond traditional comparative analysis, placing shotgun metagenomic data in the context of community-level metabolic networks. Comparing the topological properties of the enzymes in these networks with their abundances in different metagenomic samples, and examining systems-level topological features of microbiomes associated with different host states allows us to obtain novel insight into variation in metabolic capacity. This approach extends the metagenomic gene-centric view by taking into account not only the set of genes present in a microbiome, but also the complex web of

interactions between these genes and by treating the microbiome as a single 'independent' biological system [65].

Computational systems biology methods and complex network analyses have been applied widely to study microorganisms, and a variety of approaches have been developed to create genome-scale metabolic networks of various microbial species [66]–[68]. In this study, we focus on simple connectivity-centered networks that are computationally derived from homology-based large-scale metabolic databases [69] coupled with a topological analysis. These networks form a simplification of the actual underlying metabolic pathways and may be relatively inaccurate and noisy. Yet, topology-based analysis of such networks has proved powerful for studying the characteristics of single-species metabolic networks and their impact on various functional and evolutionary properties, including scaling [70], metabolic functionality and regulation [49], [71], modularity [51], [72], essentiality and mutant viability [73], genetic and environmental robustness [50], adaptation [74], [75], and species interaction [76]. To date, however, topological analysis has not been used to examine community-level metabolic networks and to study metagenome-scale metabolism.

.

2.3 Methods

2.3.1 Datasets

Metagenomic data was obtained from two studies of the human gut microbiome. The first study [16] examined 576.7 gigabases of Illumina-derived sequences from 124 European individuals labeled with BMI (kg/m^2) and inflammatory bowel disease (IBD) data. The second study [43] examined 454 FLX–derived sequences from 6 twin-mother trios from the Missouri Adolescent Female Twin Study (MOAFTS) binned according to BMI. All sequence

data was mapped to KEGG orthologous groups (KOs) using BLASTX (see Appendix A for additional details).

2.3.2 Enzyme Enrichment

To identify enzymes (KOs) that are associated with obesity, the abundance of each enzyme in the set of samples obtained from obese individuals was compared with its abundance in lean/overweight individuals. To prevent the confounding effects of overlapping host states, samples labeled with IBD were excluded from this analysis. For each enzyme, *k*, an odds ratio was calculated according to $OR(k) = [\sum_{s=obese} A_{sk} / \sum_{s=obese} (\sum_{i \neq k} A_{si})] / [\sum_{s=lean} A_{sk} / \sum_{s=lean} (\sum_{i \neq k} A_{si})]$ where A_{sk} denotes the abundance of enzyme *k* in sample *s, 'obese'* denotes the set of obese samples, and *'lean'* denotes the set of lean/overweight samples (Fig. A.3). See SI for more details on this choice of enrichment metric. The differential abundance score was defined as the absolute value of the fold change in odds ratio, $abs(\log_2(OR))$. Obesity-associated enzymes were those with a differential abundance score >1. Obesity-associated enzymes were further classified as obesity-enriched (OR>2) or obesity-depleted (OR<0.5). IBD-associated enzymes were identified in a similar manner. When calculating IBD-associated odds ratios, samples labeled as obese were excluded from the analysis. A more stringent odds ratio-based analysis was used to identify enzymes that were *consistently* enriched or depleted (Appendix A). Additionally, a number of other statistical methods were used to quantify differential abundance and identified enzymes associated with a given host state (Appendix A). Repeating the analysis with these alternative methods yielded qualitatively similar results (Appendix A and Table A.1).

2.3.3 Network Construction

A community-level metabolic network was constructed from the entire set of enzymes found in any sample (see SI). The KEGG database was used to annotate enzymes with metabolic reactions. Each enzyme may be associated with multiple reactions,

and each reaction may be associated with multiple enzymes. Using this mapping, an enzyme-based metabolic network was constructed, where nodes represent enzymes (KOs) and a directed edge from enzyme 1 to enzyme 2 indicates that a product metabolite of a reaction catalyzed by enzyme 1 is a substrate metabolite of a reaction catalyzed by enzyme 2 (Fig. A.1B). For both datasets, 98% of the enzymes were part of a single giant connected component. The network was trimmed to include only the nodes and edges that were part of this giant component, and only these enzymes were used in the subsequent analysis. To create host state specific networks, the same procedure was followed using only the set of enzymes recovered from samples in a given host state.

2.3.4 Topology-Based Measures and Analysis

Topological features of each enzyme in the network were calculated with the Cytoscape NetworkAnalyzer plugin [77]. The overall correlation between all topological features supported by the NetworkAnalyzer plugin was calculated and a feature set without any pairwise correlations >0.95 was selected for further analysis. This feature set included betweenness centrality (defined as the proportion of shortest paths passing through a node), clustering coefficient (defined as the proportion of existing edges between a node's neighbors), neighborhood connectivity (average number of neighbors of a node's neighbors), indegree (the number of edges terminating in a node), and outdegree (the number of edges originating in a node). Fig. A.2B provides additional illustrations and examples of these features (and see also: *http://med.bioinf.mpi-inf.mpg.de/netanalyzer/help/2.6.1/index.html*). The *betweenness centrality* feature was used throughout the paper to measure the centrality of each enzyme in the network. Enzymes were further classified as *peripheral*, *intermediate*, or *central* by ranking all enzymes according to centrality and partitioning this ranked list into three equally populated bins, which we termed centrality tiers.

The Spearman correlation test was used to examine the correlation between differential abundance scores and each topologic feature. A Wilcoxon rank-sum test was used to compare the topology scores of host state associated enzymes (and specifically enriched or depleted enzymes) to the scores obtained for non-differentially abundant enzymes. A Hypergeometric enrichment test was used to examine the over-representation of host-state associated enzymes in each centrality tier.

2.3.5 Network-Level Topological Features of Host State-Specific Networks

Samples were divided into three distinct groups: lean-healthy, obese-healthy, and lean-IBD. The three obese-IBD samples were not used in this analysis. Three separate host-state specific networks were created from the pooled set of enzymes identified within each group. Network-level features including node count, density (the ratio of edges to nodes), and modularity were calculated for each network. Here, we define and calculate modularity according to the formulation presented in [78]. For a particular division of a network into discrete modules, modularity is defined as the number of edges between nodes that belong to the same module minus the expected number of such edges in an equivalent randomized network, normalized by the total number of edges. The modularity of the network is calculated for the division that maximizes this value (see [78] for a complete mathematical formulation). This modularity value measures how well a network can be partitioned into densely connected modules with relatively few edges running between modules. Rarefaction curves were generated for each of these measures by considering an increasingly larger random subset of reads from each group. The statistical significance of these measures was assessed compared to null distributions calculated from randomized networks (SI).

2.3.6 Seed-Set Identification.

The metabolic seed set (see SI for more details), representing enzymes operating on exogenously acquired compounds, was calculated according to the method described in [75].

2.4 Results

2.4.1 Datasets

Illumina-derived shotgun metagenomic data from 124 unrelated Danish and Spanish individuals were analyzed [16]. Of the 124 individuals, 82 were labeled as lean/overweight (BMI<30) and 42 were labeled as obese (BMI≥30). Additionally, 25 were diagnosed with inflammatory bowel disease (IBD) relative to 99 healthy individuals. IBD patients were all of Spanish descent, and Spanish individuals were mostly labeled as lean. An additional dataset, comprising 454 FLX–derived data from six obese and lean monozygotic twin pairs and their mothers [43], was analyzed as well. When applicable, we applied our analysis to this second independent dataset to confirm the validity of our results (SI). See Materials and Methods for a detailed description of each dataset.

2.4.2 Obtaining Community-Level Metabolic Networks

To construct a community-level metabolic network of the gut microbiome, metagenomic sequence reads were annotated using the KEGG database to identify enzymatic genes (Materials and Methods). In total, 1610 enzymes were identified and annotated with a metabolic reaction. Overall, relative enzyme abundance across the 124 samples was highly concordant (average pair-wise correlation coefficient R=0.94; Spearman correlation test), in accordance with previous studies revealing inter-sample similarity in gene content [43]. The annotation data from all samples was pooled and a network was created in which nodes represented enzymes and enzymes catalyzing successive reactions were connected by directed edges. We excluded enzymes that were not part of the largest

connected component of the network, resulting in a total of 1570 enzymes (Materials and Methods).

2.4.3 Identifying Enzymes Associated with a Given Host State

We compared the abundance of enzymatic genes across various samples to identify enzymes associated with a given host state (e.g. obesity or IBD). Specifically, we used an odds ratio (OR) test to measure the fold change in the abundance of an enzyme in samples taken from hosts with the given state compared to its abundance in other healthy samples (Materials and Methods; and see SI for further details on the metric choice). The differential abundance score of each enzyme, defined as abs(log2(OR)), provides a measure of the extent to which an enzyme's abundance differs in samples from a given host state, relative to healthy samples. Enzymes with differential abundance score higher than 1 (i.e., enzymes that are either two-fold enriched or two-fold depleted) are defined as being associated with the given host state.

To verify that the results reported below are not dependent on the specific choice of enrichment metric used, we further examined several alternative methods for identifying host state-associated enzymes (including significance analysis, presence/absence over-representation test, rank-based difference test, and distribution divergence analysis; see SI for more details). These enrichment metrics yielded qualitatively similar results (Appendix A and Table A.1). Similarly, to confirm that our findings do not stem from potential noise in the read count data, we used a shuffling analysis to identify enzymes that are 'consistently' enriched or depleted across samples (Appendix A). Using this more stringent criterion for enzymes associated with a given host state did not qualitatively change the results below (Appendix A and Table A.1).

An over-representation analysis (Appendix A) showed that enzymes that are enriched in obese or IBD microbiomes are more frequently involved in membrane transport

($p<0.035$ [obese]; $p<0.006$ [IBD]; Table A.2). These results are consistent with previous analysis of enriched functions in the smaller dataset of lean and obese twins [43]. In contrast, enzymes that are depleted in obese microbiomes are more frequently involved in cofactors and vitamins metabolism ($p<0.03$), nucleotide metabolism ($p<0.002$), and transcription ($p<2.52\times10^{-12}$) amongst other processes (Table A.2).

2.4.4 Linking Host State Associated Enzymes to Centrality

Using the community-level network outlined above, we examined whether enzymes that are associated with a specific host state exhibit unique topological features. We first focused on a topologically-derived centrality measure termed *betweenness centrality* [49]. This measure calculates the proportion of shortest paths in a complex network that pass through a given node, as a proxy for the node's location in relation to all other nodes (Fig. A.2B). High centrality values are typically associated with nodes located in the core of the network whereas low centrality values indicate a more peripheral location.

We found that an enzyme's differential abundance score in obese samples is negatively correlated with its centrality in the network ($R=-0.17$, $p<1.3\times10^{-12}$; Spearman correlation test). Partitioning the set of enzymes in the network into those that are associated with obesity (as defined above) and all other enzymes, we similarly find that centrality scores of obesity-associated enzymes are significantly lower ($p<8.9\times10^{-6}$; Wilcoxon rank-sum test; Fig. 2.1A). As further validation, we note that decreased centrality is not associated with equivalent sets of randomly selected enzymes ($p<8\times10^{-4}$). Significantly lower centrality scores can also be observed when examining obesity-enriched and obesity-depleted enzymes separately ($p<0.03$ and $p<7.4\times10^{-6}$ respectively; Wilcoxon rank-sum test; Fig. 2.1A), suggesting that obesity is characterized by both gain and loss of peripheral enzymes. Similarly, partitioning the enzymes in the network into three centrality-based tiers (Materials and Methods), we find a significant over-representation of obesity-associated

enzymes in the peripheral tier of the network: 29.1% of the enzymes in this tier are associated with obesity compared to only 19.4% and 18.6% of the enzymes in the intermediate and central tiers respectively (Fig. 2.1*B*). Using the more stringent criterion defined above for identifying enzymes that are consistently associated with obesity yields a similar trend; 13.6%, 10.4%, and 9.8% of the enzymes in the periphery, intermediate and central tiers respectively are consistently associated with obesity (see SI). This negative association between obesity-associated differential abundance and centrality was confirmed in the analysis of the smaller twin-mother trios dataset ($R=-0.15$, $p<9.7\times10^{-8}$; see SI for additional results).

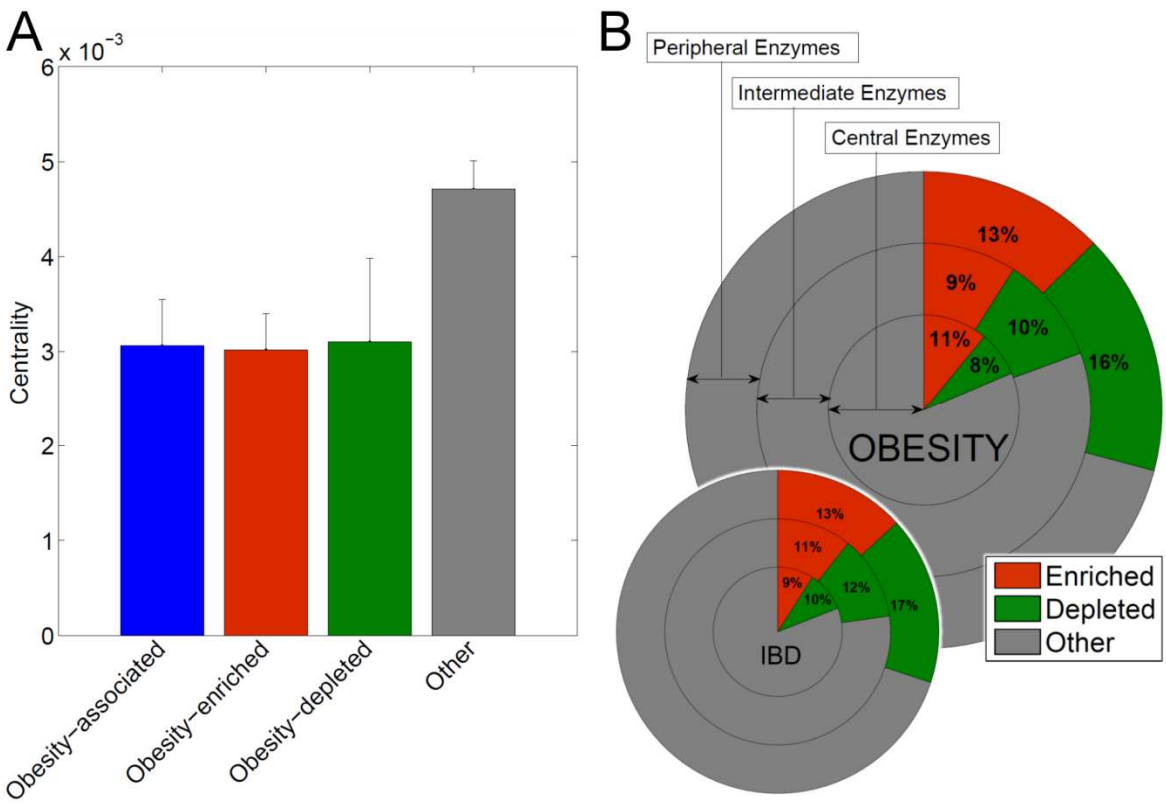


Figure 2.1 Network centrality of host-state associated enzymes (A) Mean and standard error of the centrality scores of obesity-associated enzymes vs. all other enzymes in the network. Obesity-associated enzymes are further divided into enzymes that are enriched or depleted in obese microbiomes. (B) Proportion of enzymes that are associated with obesity (main plot) and IBD (inset) within three equally populated centrality-based network tiers. Each concentric pie chart depicts the percent of enzymes within a specific centrality tier that are classified as enriched or depleted. Enzymes associated with obesity or IBD are found in

significantly higher proportions in the peripheral tier ($p<5.6x10-6$ [obesity], $p<4.8x10-5$[IBD]; Hypergeometric enrichment test). This result still holds considering alternative or stricter criteria for association with the host state (SI).

Interestingly, a similar pattern is observed in enzymes associated with IBD. An enzyme's differential abundance score in IBD is negatively correlated with its centrality ($R=-0.15$, $p<1.9x10^{-9}$; Spearman correlation test) and the centrality scores of IBD-associated enzymes are significantly lower than the centrality scores of enzymes not associated with IBD ($p<9.5x10^{-6}$; Wilcoxon rank-sum test; $p<0.003$ and $p<0.0002$ for IBD-enriched and IBD-depleted enzymes respectively). Similarly, IBD-associated enzymes are significantly over-represented in the peripheral tier of the network: 30.1% of the enzymes in this tier are associated with IBD compared to only 22.8% and 19.0% of the enzymes in the intermediate and central tiers respectively (Fig. 2.1*B*, inset). A similar trend is observed when considering only consistently associated enzymes (8.8%, 5.4%, and 6.1% respectively).

We confirmed that the above patterns, linking host state associated enzymes to centrality, are robust to several alternative network construction methods (e.g., using the SEED annotation framework [79] rather than KEGG) and are not affected by using different threshold values to filter out low count reads and potential noise (Appendix A and Table A.1). To also validate that the above results are not the outcome of population substructure, we repeated the analysis for obesity-associated differential abundance using only the Danish individuals, and the analysis for IBD-associated differential abundance using only the Spanish individuals. Using these subpopulation samples, we still observed a significant correlation between centrality and differential abundance (Table A.1). We further confirmed that this correlation between differential abundance and centrality is not solely a product of the over-representation of transport enzymes (which are likely to be found at the periphery of the network) in obese microbiomes (Appendix A and Table A.1).

Large-scale metabolic data (such as KEGG) are often based on automated, comparison-based, genome annotation [69] and are therefore bound to be incomplete and imprecise [80]. Such inaccurate metabolic annotations may markedly affect various complex network properties and can potentially dramatically impact our results. However, using a sensitivity analysis to examine the effect of missing or erroneous annotation data (Appendix A and Figs. A.4 and A.5), we verified that the calculated centrality scores and the pertaining results reported above are fairly robust to such inaccuracies in the raw metabolic annotations.

2.4.5 Linking Host State Associate Enzymes to Additional Topological Features

We next examined a number of additional topological measures for each enzyme in the network, including indegree, outdegree, neighborhood connectivity, and clustering coefficient (Materials and Methods). In contrast to centrality, these measures are more local in nature, taking into account only the immediate neighborhood of each enzyme, and hence capture a different aspect of network topology. The seed set of the network was also identified using a previously published seed detection method [75], and consisted of 126 enzymes. The seed detection method applies a graph-theory based algorithm to analyze the topology of a given network and identify the minimal set of topological “input” nodes sufficient to activate all other nodes in the network (see Appendix A for more details). The seed sets of metabolite-based networks of a large array of microbial species were shown to be a successful proxy for the biochemical environments of these species and to provide insights into their ecology [75], [76], [81].

While both enriched and depleted enzymes exhibit low centrality as described above, we found that enriched enzymes differ dramatically from depleted enzymes in respect to such local topological features. Specifically, enzymes enriched in obese microbiomes have

a significantly lower clustering coefficient ($p<7.8\times10^{-4}$; Wilcoxon rank-sum test) and lower indegree ($p<0.004$) compared to enzymes that are not associated with obesity (Fig. 2.2*A*-*B*). In contrast, enzymes depleted in obese microbiomes have a significantly higher clustering coefficient ($p<0.006$; Wilcoxon rank-sum test) and higher indegree ($p<0.02$) compared to non-associated enzymes. IBD-associated enzymes follow similar trends but are not statistically significant due to smaller sample size (Fig. A.6). We additionally found that enzymes identified as network seeds have significantly higher differential abundance scores ($p<4.8\times10^{-6}$; Wilcoxon rank-sum test) compared to non-seeds and that such network seeds are over-represented amongst obesity- and IBD-associated enzymes ($p<2.7\times10^{-4}$ [obesity] and $p<2.6\times10^{-3}$ [IBD]; See SI for more details).

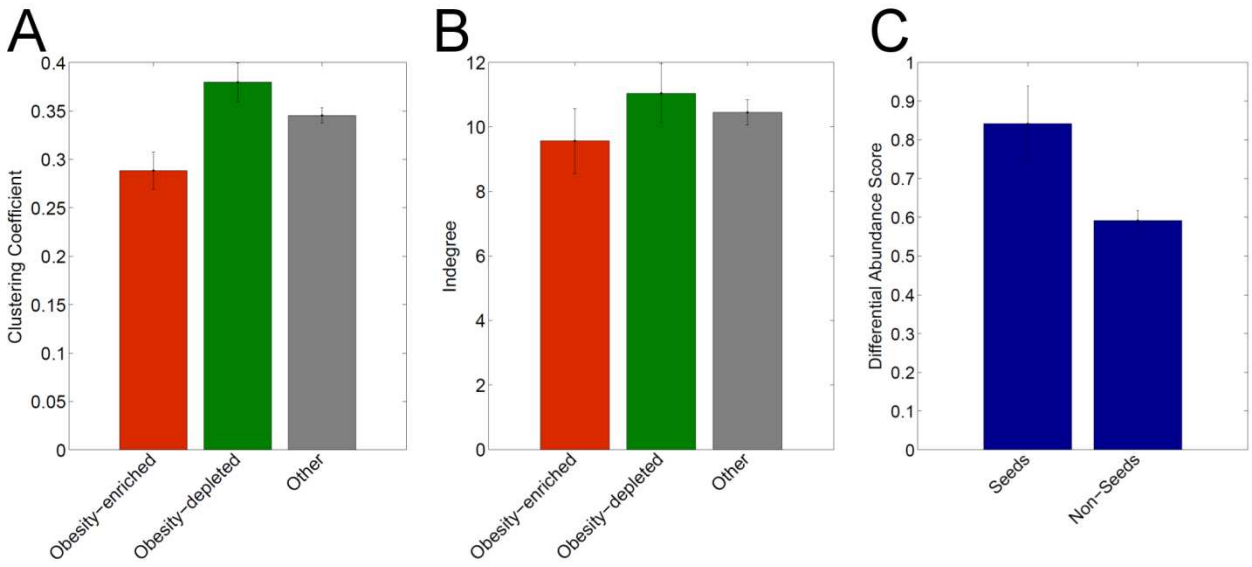


Fig. 2.2. Host-state associated shifts in local network topology. Mean and standard error of the clustering coefficient **(A)** and indegree **(B)** of enriched (red; n=170), depleted (green; n=180), and other (gray; n=1213) enzymes in obese microbiomes. Clustering coefficient is defined as the ratio between the total number of edges connecting a node's neighbors and the potential number of edges that could exist between them. Indegree denotes the number of edges terminating at a node. **(C)** Mean and standard error of the differential abundance scores of seeds vs. non seed enzymes.

Such distinct topological properties may additionally be used as potentially informative attributes and to highlight biomarkers for involvement in obesity and IBD.

Specifically, we examined enzymes enriched in obese or IBD microbiomes and within these sets focused on enzymes that also exhibit the topological features identified above (low centrality, low indegree, and low clustering coefficient). We find that a large fraction of these enzymes within both the obesity-enriched or the IBD-enriched enzymes are involved in either the phosphotransferase system (PTS) (28.6% amongst obesity-enriched enzymes and 20.6% amongst IBD-enriched enzymes), or the nitrate reductase pathway (17.1% and 17.6% amongst obesity- and IBD-enriched enzymes respectively). Notably, PTS is a Eubacteria-specific strategy for transporting sugar into the cell, and has been specifically associated with members of the Firmicutes phylum [82]. Use of this transport system has been implicated in regulation of carbohydrate uptake [83], and was found to be upregulated following a switch to a high-fat/high-sugar 'Western' diet in mice [84]. Recently a PTS enzyme (FrvX) was found to be a biomarker for IBD [85]. Similarly, nitrate reductase is a critical component in the conversion of nitrate into nitrite and nitric oxide, and is not synthesized by human DNA. Elevated levels of nitric oxide have been associated with both IBD [86] and obesity-induced insulin resistance [87], as well as other serious carcinogenic and inflammatory effects [88]. We additionally find in this set enzymes for xenobiotic metabolism, most notably those for the metabolism of choline and p-cresol, which have been linked to various host diseases and metabolic phenotypes (see SI).

2.4.6 Linking Topological Variation to Community Species Composition

Shotgun metagenomic data and community-level models provide a functional view of community metabolism. Ultimately however, differences in community gene content reflect differences in species composition. Understanding the link between variation in community-level topological properties and community composition can provide valuable insight into the mechanism by which community activity changes as a result of compositional shifts. As a full decomposition of shotgun metagenomic data into species-specific data is not

yet feasible, we studied the distribution of genes of interest across a large array of reference genomes. Specifically, examining the genomes of 326 fully-sequenced prevalent gut-dwelling microbial species (Materials and Methods), we found that enzymes associated with either obese or IBD microbiomes tend to be present in fewer genomes than non-associated enzymes ($p<10^{-54}$ [obesity], $p<10^{-56}$ [IBD] Wilcoxon rank-sum test; Fig. A.7). Obesity-associated enzymes were also present in fewer genomes than randomly selected sets of enzymes ($p<10^{-4}$; SI). Moreover, the centrality of enzymes in the community-level metabolic network is correlated with the number of reference genomes in which these enzymes occur ($R=0.23$, $p<10^{-17}$; Spearman correlation test; Fig. A.8). A universal association between centrality and prevalence has also been recently demonstrated for a smaller set of species that were not associated with the human microbiome [89]. These findings suggest that the variation in community-level metabolism associated with obesity and IBD may be induced by an increase or decrease in the abundance of a relatively small subset of species.

2.4.7 Linking Host State to Network-Level Topological Properties

Finally, we examined whether host state-associated differences also translate into differences in network-level topological features. Sequence reads derived from lean-healthy, obese-healthy, and lean-IBD samples were pooled separately and used to construct state-specific metabolic networks. Calculating various network-level topological features for each of these networks, we found that the variation associated with host state goes beyond a limited set of enriched or depleted enzymes and also induces global differences in network topology. Specifically, obese microbiomes were found to induce a less modular metabolic network than lean microbiomes. Interestingly, reduced modularity in the metabolic networks of single species has recently been associated with lower variation in the environment (see also Discussion). A rarefaction analysis was performed to confirm that all networks derived from each of the three sample groups reached a stable topology within the available

coverage (Fig. 2.3*A*). An extensive shuffling-based analysis (Fig. 2.3*B*, Appendix A, Figs. A.9-A.11) demonstrated that the difference in the level of modularity between obese and lean microbiomes is statistically significant ($p<0.027$) and is not expected at random from multiple individual realizations of networks with similar topological properties.

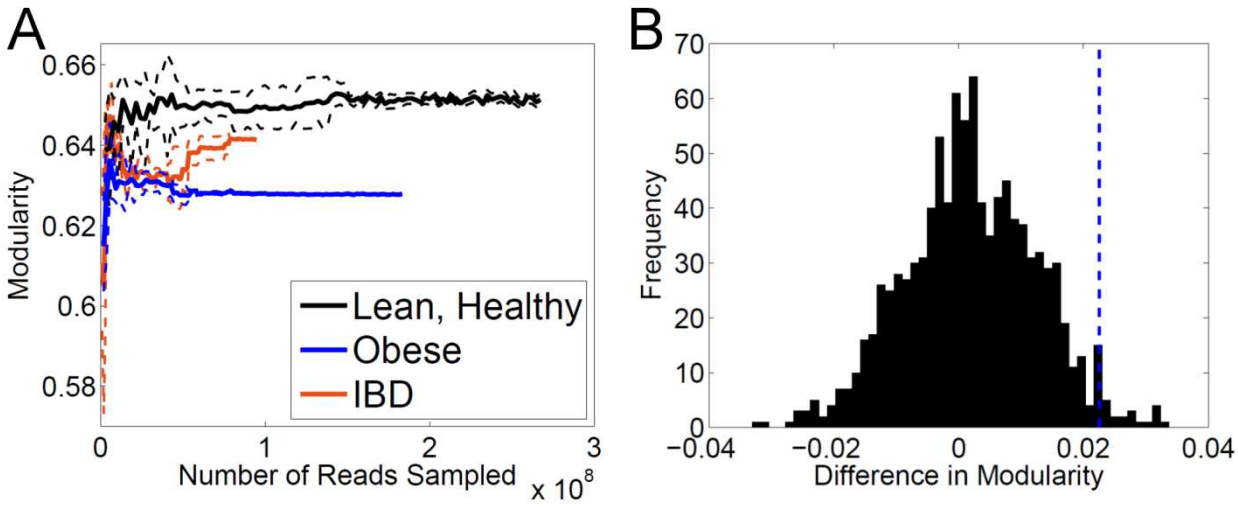


Fig. 2.3. **Modularity of host-state-specific metabolic networks. (A)** Rarefaction analysis of the modularity of pooled lean-healthy, obese-healthy and IBD-lean microbiomes. The plot depicts the mean (solid line) and standard deviation (dotted lines) of 5 rounds of rarefaction analysis, obtained by calculating the modularity of networks derived from progressively smaller randomly selected sets of reads. (B) The difference between the modularity of the obese-specific and lean-healthy-specific network is plotted (dashed blue line) against a null distribution of differences obtained via random grouping of samples (see SI for more details). The observed difference in modularity is significantly greater than the expected difference according to this null distribution.

2.5 Discussion

Taken together, the topological features that were found to vary with obesity and IBD suggest a characteristic mode of deviation from a normal microbiome organization that may be associated with a disease state. This suggests that in addition to, or potentially as a consequence of alterations in the abundance of individual genes or functional classes, disease may be associated with higher-order modes of deviation in the microbiome. Clearly, such associations alone cannot directly implicate a mechanism for disease; both obesity and

IBD are poorly understood diseases and embody extremely complex phenotypes. Accordingly, the system-level observations reported in this study can have multiple alternative interpretations and stem from mechanisms that are yet unknown. These observations, however, offer a number of possible interpretations and allow us to posit intriguing hypotheses for further study.

Specifically, we find that enzymes that typify various host states tend to have low centrality and are found mostly in the periphery of the network. As the topology of the network reflects metabolic interdependencies between enzymes (rather than physical location in the gut) the periphery of the network represents metabolic steps that are relatively remote (as measured by their distance along various metabolic pathways) from the core of the network and that are closer functionally to the microbiome environment [90]. The most peripheral enzymes, for example, represent either the microbiome's first metabolic steps (i.e., enzymes that rely on substrates that are not produced by any other enzyme in the microbiome) or end points (enzymes that produce metabolites which are not utilized by other microbiome enzymes). Such enzymes are likely to directly use or produce metabolites that characterize the gut environment, forming an interface between microbial and human metabolism. Our results therefore suggest that much of the enzyme-level variation associated with obesity or IBD relates to changes in the way the microbiome interacts with the gut environment rather than variation in core metabolic processes. This variation corresponds to both gain and loss of certain peripheral metabolic enzymes, as suggested by the reduced centrality of both enriched and depleted enzymes. This is also supported by the reported link between differentially abundant enzymes and seed enzymes. Obesity-enriched enzymes, however, specifically possess further topological properties characteristic of network input points (low indegree and low clustering coefficient). While several mechanisms that link the microbiota to obesity have been reported, this finding may suggest

that obese microbiomes are capable of utilizing a diverse repertoire of energy sources, accounting for their increased capacity for energy extraction from the diet [40]. Interestingly, it has also been shown that functionally peripheral enzymes (those involved in nutrient uptake and first metabolic steps) are more likely to be horizontally transferred [90] and are gained and lost more frequently during the evolution of individual microbial organisms [75]. This similarity between the adaptive variation that occurs in single species across an evolutionary time scale and community-level variation across samples further supports our treatment of the community as a comprehensive biological system.

Our topology-based system approach has further suggested candidate biomarkers involved in obesity and IBD. In addition to phosphotransferase systems used for the import of dietary carbohydrates, both obesity and IBD were significantly associated with genes for the production of NO_2 and the metabolism of choline and p-cresol. The unexpectedly high overlap between these disease-associated gene sets may be indicative of some common underlying triggers of disease or alternatively a conserved response of the gut microbiome to disease. Follow-up studies using gnotobiotic mouse models colonized by microbial isolates with the ability to perform these key functions, “humanized” mouse models colonized with samples taken from paired healthy and diseased human donors, and human intervention studies will be critical to determine which aspects of the gut microbiome may contribute to disease and the precise mechanisms that link this complex microbial metabolic network to host physiology.

Our results further demonstrate that the variation associated with obesity and IBD induces a reduced network-wide modularity. Recent studies of metabolic network topologies across the bacterial tree of life revealed marked variation in network modularity and identified several genetic and environmental determinants affecting metabolic modularity [51], [91]. Specifically, these studies demonstrated that reduced metabolic modularity in

single species networks is associated with organisms inhabiting less variable environments. Our analysis, however, presents the first characterization of community-level modularity and demonstrates consistent differences that are associated with the host state. It is intriguing to extrapolate findings from single species analyses and to hypothesize that reduced community-level modularity in obese microbiomes may be associated with a decreased variability in the gut environment or with the lack of temporal regularities [92]. Furthermore, this reduced modularity may be construed as a functional manifestation of the reported decrease in species diversity that has been observed in obese individuals [43].

In silico models of microbial communities are currently still scarce [66] and mostly focus on simulated communities comprising a handful of species and on pair-wise interactions between community members [13], [68], [93], [94]. Here, in contrast, we take an integrative approach, treating the microbiome as a single, supra-organismal system [64], and examining the metabolic network of the community as a whole [95]. Moreover, this study is the first to focus on the topology of this metagenome-based network and on the relationship between its topology and the host state. As with any attempt to represent a dynamic and stochastic set of biological processes via a simple model, our analysis is subject to various assumptions and simplifications. Our framework ignores boundaries between species and compartmentalization of various metabolic processes (see also SI on analyzing communities as supra-organisms). Additionally, as mentioned above, topological analysis of connectivity-based and static networks explicitly ignores several features of metabolic reactions such as metabolic rates and dynamic regulation. Furthermore, our analysis considers metabolism alone and does not account for other processes that may be involved (e.g., immune response). Such simple models, however, are extremely useful for studying systems for which data is still limited and our ability to construct more involved models is hindered. Here, for example, they facilitate the integration of multiple modes of

microbiome characterization, and support analysis using the rich set of tools developed for systems biology and complex network analysis. As our understanding of the human microbiome improves, better models can be constructed, potentially using the collective effort of the research community [96], [97]. Experimental validation of model components and parameters is crucial for a successful and accurate reconstruction. Moving forward, microbiome-wide models can further integrate transcriptional and metabolomics-based data. Such manually curated models may ultimately provide a predictive framework, similar to the one available for individual species, for targeted community manipulation and for informing clinical interventions.

In essence though, this study represents an important step in the construction of a novel 'metagenomic systems biology' approach. Such an approach can potentially advance metagenomic research in the same way systems biology advanced genomics; appreciating not only the parts list of a system, but also the complex interactions between parts and the impact of these interactions on function and dynamics. Future work will also include identifying specific sets of enzymes responsible for systems-level patterns, characterizing the implications of various topological variations, and linking this variation to changes in species composition. Clearly, our understanding of the complexity of the gut microbiome is still lacking and much work still remains to be done before exact mechanisms are identified. Future clinical applications may focus on specific functions rather than on system-level properties of the microbiome. Single-species metabolic models have been effective at predicting how an individual organism may function within specific culture conditions, while systems level models are required to determine the effect of the addition or deletion of whole genomes or specific genes on the entire metabolic system. Predictive studies for example, may focus on identifying the suite of functions that modulate the effect of the addition of a probiotic species or the ramifications of antibiotic treatment on community function. Meanwhile, pinpointing a certain

section or link missing from an individual's microbial metabolic map could point doctors to more targeted treatments, obviating the need for highly disruptive broad-spectrum antibiotics. In all, this systems biology approach provides a complementary viewpoint to comparative and functional metagenomics in gaining valuable intuition concerning the function of the microbiome as a system and in identifying potential biomarkers for further validation.

3. Assembly Modules of the Mammalian Microbiome

3.1 Summary

Mammalian microbiomes differ in species and gene composition, and these differences have been shown to correlate better with host diet than host phylogeny. However, the processes that underlie this variation are unclear, specifically how gene relationships contribute to structure in metagenome composition. Here, we present a method for detecting metagenomic assembly modules, sets of genes that are both functionally related, and co-occur across microbiomes. Applying this method to a panel of 39 mammalian microbiome samples, we detect 29 assembly modules, some of which comprise gene sets that may be contributed by multiple microbial species. Importantly, we find that variation in module representation across individual samples gives deeper insight into how host diet shapes microbiome composition, and sheds light on the functional processes that comprise each mammal's microbial metabolic system.

3.2 Background

The diverse assemblages of microbial species and genes inhabiting the mammalian gut (the mammalian gut microbiome) have been shown to be important for a number of metabolic and immune-related processes. The composition of this microbiome can vary widely across hosts. The full spectrum of forces responsible for producing such compositional variation are still unclear, however recent studies ([98], [99]) have examined associations between the composition of the mammalian gut microbiome and host evolution or host diet. Interestingly, overall microbiome species composition was found to be more closely associated with host lifestyle (ie. diet) than with phylogeny, suggesting that the mammalian microbiome is shaped primarily by environmental adaptation rather than vertical transmission. Further analysis [100] revealed that while mammalian hosts shared a large functional core, gene composition reflected

species composition and was associated with host diet as well. 495 genes were identified that differed significantly in relative abundance between host diet groups, while by contrast, only 18 genes were identified that were significantly associated with host phylogeny.

These investigations are essentially asking about the processes that shape microbiome assembly. The results suggest that the microbiomes of phylogenetically diverse hosts converge during adaptation in response to similar biochemical gut resources. However, what remains unclear are the functional units of this adaptation. Does the presence or abundance of an individual gene fundamentally change the processes a microbiome is capable of achieving? Or, as suggested by studies of metabolism in single species[101], are the metabolic capabilities of a microbiome defined by the collective presence or abundance of coherent groups of genes, all relevant for a given functional outcome? While the previous studies show that individual genes differ between host diet groups, and consequently overall gene profiles differ, functional relationships between genes are not accounted. Rather, each gene is treated as a separate entity, obscuring the extent to which the *capabilities* of any two microbiomes differ.

Addressing such questions instead calls for a systems-level approach. Integrating metagenome composition with known gene roles and relationships would allow us to better understand the components that comprise microbiome assembly and the potentially discrete ways in which composition may vary. We would expect that if a particular metabolic task was accomplished by a defined set of interacting genes, then when looking across highly diverged microbiomes – for example those from different mammalian species – we would find certain host species for whom this task was pertinent and in which the entire gene set was present, and other host species for whom this task was irrelevant, and in which none of these genes appear with frequency. Identifying such sets of co-occurring genes that also have known *functional* relationships would both provide support for the biological validity of the detection, and give insight into why such a set may appear in certain host species, and what adaptive forces govern

its presence or absence. Though similar in concept to pathways and co-evolutionary modules that have been previously defined by gene co-occurrence patterns across microbial genomes[102], [103], such gene sets have not previously been defined for *metagenomes*. We term these sets *metagenomic assembly modules*, and set out to identify these modules in the gut microbiomes of a diverse panel of mammalian host species.

We start by demonstrating that the previous findings regarding diet-associated adaptation are echoed in systems-level features of a microbiome metabolic model, and that our modeling framework captures certain salient features of host diet that are reflected in microbiome composition more efficiently than overall gene composition profiles. We then move on to the detection of assembly modules, using a simple greedy algorithm to identify coherent groups of co-occurring, functionally-related genes. We hypothesize that each mammalian microbiome is composed of a unique set of modules, and that certain modules are relevant for adaptation to host diet, while others may represent host species-specific, or even organism-specific adaptation.

3.3 Methods

3.3.1 Mammalian metagenomic samples and KO abundance

Raw short-read sequences (150-400bp) from 39 mammalian gut microbiome samples (21 herbivores, 7 omnivores, 11 carnivores) were downloaded from MG-RAST[54] (project ID 116) and rarefied by random selection of reads to the read count of the sample with the lowest read depth (Lion1; 14,236 reads). Samples were then annotated with KEGG orthologous groups (KOs) via an in-house pipeline[104] and the KEGG database (version downloaded 07.15.2013), yielding matches to 5,428 KOs across the full sample set. KO counts were normalized and corrected via the MUSicC pipeline, to obtain per-genome KO copy number averages (here, referred to simply as *KO prevalence*). KO presence was in a sample was defined as KO prevalence >0.1(Fig. B.1). At this threshold 4,904 KOs were

present in at least one sample. KOs that were differentially abundant between host diet groups were identified via two-sample t-tests, and significant KOs were those with FDR rate <.05. KOs were also mapped to KEGG pathways and KEGG functional classes to create functional profiles.

3.3.2 Metabolic Network Model Construction

Metabolic reactions associated with each of the detected KOs were obtained from the KEGG database. A pan-mammalian metabolic network model was constructed in which two KO nodes were connected with a directed edge if any of the products of a reaction catalyzed by one KO were used as substrates in a reaction catalyzed by a second KO. Reactions involving eight currency metabolites, identified as those with the highest node degree in a metabolite-based network created from the same data, were omitted during final KO-based network construction. The final network comprised 1,421 KOs. Individual networks were also similarly constructed for each sample, and ranged from 434-841 KOs.

3.3.3 Metabolic Seed Set Identification

A previously published algorithm [75] was used to identify the set of metabolic seed KOs in each sample-specific network. These are KOs operating directly on exogenously acquired metabolites and comprise the minimal set of topological inputs required to initiate every path through the network. Each seed KO is given a seed score representing its essentiality – for example, if any of 3 KOs could serve to initiate the same set of paths through the network, each KO would be given a seed score of 1/3. Seed KOs were used to generate a bi-partite network in which each host mammal node was connected to the set of nodes representing its seed KOs. An force-directed layout algorithm was used to view the resulting network and assess sample clustering in Cytoscape[105].

3.3.4 Metagenomic Assembly Module Detection

A greedy heuristic algorithm was implemented to identify modules of functionally-related KOs with high co-occurrence across mammalian samples. Two KOs were considered functionally related if a KEGG pathway existed containing both KOs, and co-occurrence was calculated from each KOs presence/absence profile across the host species. Only KOs present (KO prevalence >0.1) in at least one sample were considered. The algorithm begins by randomly selecting a *start KO* from the set of KOs with pathway annotations. The *neighbors* of this start KO are then identified – defined as the set of all KOs sharing at least one pathway annotation with the *start KO*. The jaccard distance between the presence/absence profile of the start KO and the profile of each of its neighbors is calculated to determine if any of these neighbors significantly co-occur with the start KO. Significant jaccard distances were defined as those at least 1.5 standard deviations lower than the mean jaccard distance calculated from the set of all pairs of functionally-related KOs. If a significant co-occuring neighbor exists, it is added to the start KO to form a proto-module. If more than one significant neighbor exists, the one with the lowest distance is added, and if multiple neighbors are tied, a random one is chosen. The search continues by identifying all neighbors of all KOs in the proto-module, and adding the one with the lowest mean distance from all KOs already in the proto-module, provided the mean is still significant. When no neighbors with significant mean distances can be found, the module extension search stops. Only proto-modules with at least 4 KOs were considered valid modules, and these KOs are then removed from the search pool. Once the algorithm finds 5000 unsuccessful proto-modules in a row (those with <4 KOs), the algorithm stops and the module set is complete.

Modules were manually assigned a name based on the most common functional annotations of constituent KOs. *KEGG coherency* for each module was defined as the fraction of KOs in the module annotated with the most common KEGG pathway, divided by

the total number of KEGG pathways represented. The representation of each module in each sample was defined as the fraction of module KOs present in the sample. A two sample t-test was then used to compare module representation between host groups. The prevalence of a module was defined as the percent of samples with a representation >0.5 for that module.

3.3.5 KEGG pathway coherency

KOs in each module were associated with one or more KEGG pathways. Within each module, the relative contribution of each pathway was determined as a weighted sum of KO pathway membership, in which KOs that were part of multiple pathways contributed equivalent fractional counts to each of these pathways. Pathway coherency for each module was defined as the mean relative contribution across all pathways represented in the module. For example, if in a module of 8 KOs, all KOs belonged to only one identical pathway, the coherency would be 1. If half of the KOs also belonged to a second shared pathway, 4 KOs would contribute fully to pw1, the other 4 would contribute a total count of 2 to pw1 and 2 to pw2. The coherency would then be the mean of 6/8 and 2/8 = 0.5.

3.3.6 *Phylogenetic coherency*

KO copy number profiles for 2,337 sequenced microbial reference genomes were downloaded from IMG[55] and converted to presence/absence profiles for each KO detected in the pan-mammalian microbiome. The phylogenetic coherency of KO pairs within modules was compared to that of all functionally-related KO pairs by calculating 1 - the jaccard distance between each pair of relevant profiles.

3.4 Results

3.4.1 Metabolic seed set discriminates host diet groups

Reads from each mammalian microbiome were annotated with KEGG orthologous groups via BLASTx and normalized to obtain the prevalence of each KO across the

microbial genomes in each sample. KOs detected at a prevalence >0.1 were considered present in a sample. Overall, this yielded 4,904 KOs present in at least one of the 39 samples. Our new annotation process yielded qualitatively similar results to those reported in the original study. Specifically, a principal component analysis of overall KO profiles loosely separated mammals in different diet groups (Fig. B.2), while the membership of the KOs detected for each sample in different KEGG pathways and broader functional classes was relatively stable (Fig. B.3). Using t-tests to define diet-associated KOs within a pan-mammalian metabolic network model, we identified 318 KOs with significantly differential abundance between carnivores and herbivores. Approximately a third of these KOs (n=112) were associated with one or more of the 495 E.C.s found to have significantly different relative abundance between carnivore and herbivore microbiomes in the original study [100].

Though overall KO profiles differed between host diet groups, it is unclear which specific aspects of metabolism reflect differences in available gut resources. We reasoned that the types of KOs most likely to exhibit variation as an adaptation to diet would be those most directly responsible for harvesting different nutrients from the gut environment. Therefore, we constructed network models of the microbial metabolism of each host and from these networks, identified each microbiome's metabolic seed set. Seed sets consist of KOs associated with enzymes operating directly on exogenously acquired metabolic inputs, and have been shown to accurately reflect the host environment [75]. Using these seed KOs to generate a bi-partite network, in which edges were drawn between each sample node and all of the KO nodes comprising its seed set, we find that carnivore samples cluster separately from herbivore samples (Fig. 3.1). Omnivore samples, however, could be found in both groups. This clustering stands in contrast to previous work in which bi-partite networks generated using all KOs identified in each mammalian sample did not exhibit any

discernable clusters [100]. Furthermore, seed KOs were far more likely to be differentially represented across host groups (hypergeometric test p-value < 4.34e-06), and seed KOs showed more variation in their presence/absence profiles than other KOs (ttest p-value < 1.4e-14).

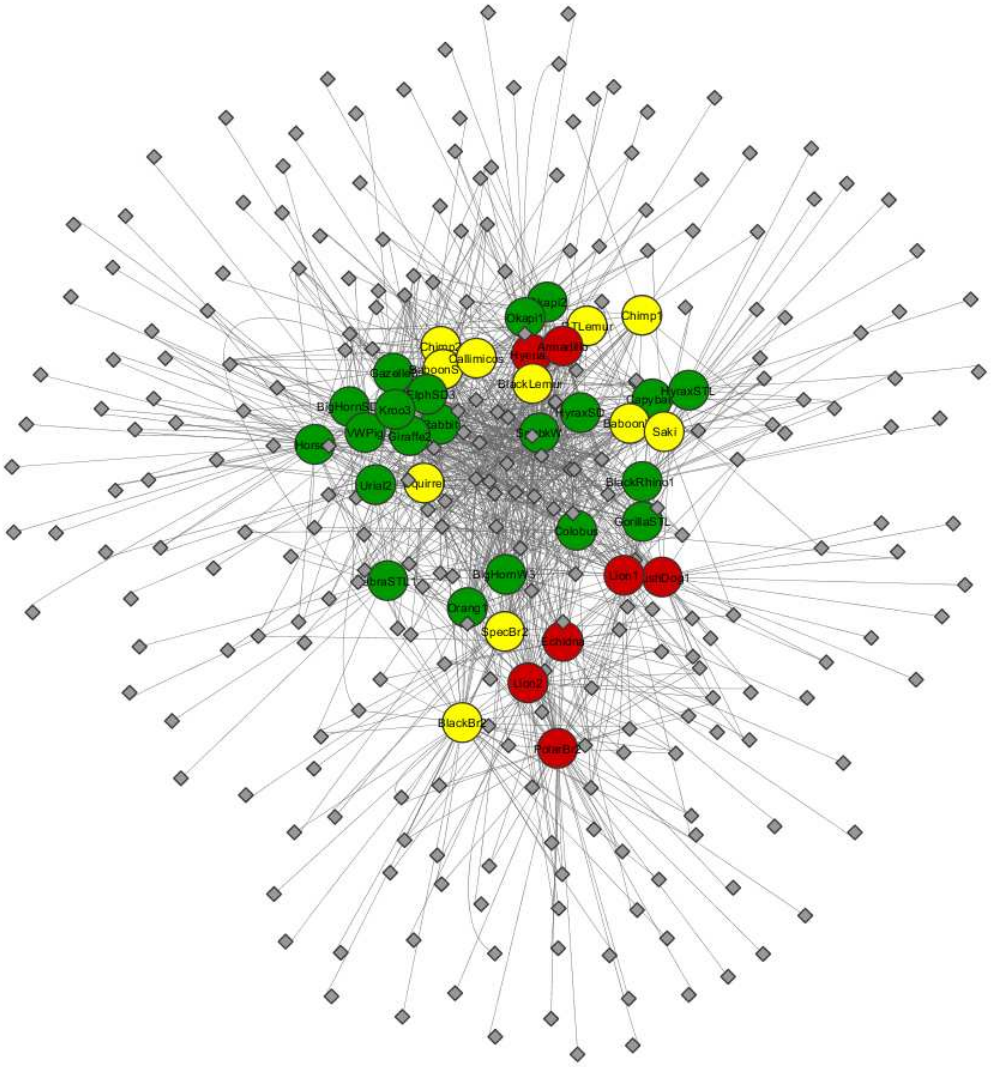


Fig. 3.1 Bi-partite network of mammalian hosts and metabolic seeds. Mammalian hosts (circles colored by diet; green: herbivore, red:carnivore, yellow: omnivore) are connected to their seed set (diamonds), the set of enzymes operating on metabolic inputs to the host's microbiome-wide metabolic model.

3.4.2 Metagenomic assembly module detection identifies co-occurring sets of functionally-related genes

Next, a heuristic search algorithm was used to identify groups of KOs that were functionally connected in a network of pathway membership, and whose profile across the 39 mammalian samples was highly similar. We term such groups of functionally-related co-occurring KOs *'metagenomic assembly modules'*, as they may represent the building blocks, or minimal units, of microbiome assembly. These modules contain KOs likely to be gained or lost across host evolution as cohesive sets. 29 modules were identified by a greedy additive search algorithm (Table 3.1; see 3.3 Methods for details). As validation of this search, we show that the set of pairwise distances between KOs in the same module (mean distance = 0.197) is significantly less than the set of pairwise distances between all functionally-related KOs (mean distance = 0.834), with a p-value < 10^{-52} (Fig. B.4). Pairwise distances within each module individually were also all significant after Bonferroni correction (Table 3.1).

Table 3.1 Metagenomic assembly modules detected across 39 mammalian gut microbiomes.

Module ID	Dominant Pathway	Module Size	Module Prevalence	Mean KO Distance	Co-occurrence p-value	Phylogenetic Coherency	KEGG Pathway Coherency
1	ABC transporters	634	39	0.25	0.00E+00	0.41	0.01
2	Ribosome	51	39	0.29	0.00E+00	0.96	1.00
3	Riboflavin metabolism	5	36	0.38	4.96E-10	0.88	1.00
4	Phosphotransferase system (PTS)	12	16	0.36	3.11E-61	0.47	0.25
5	Phosphotransferase system (PTS)	6	8	0.35	7.65E-16	0.45	0.20
6	Phosphotransferase system (PTS)	8	6	0.42	2.72E-21	0.43	0.33
7	Peptidoglycan biosynthesis	4	5	0.40	5.41E-06	0.57	1.00
8	Two-component system	5	5	0.35	5.45E-11	0.60	1.00
9	Biosynthesis of siderophore group nonribosomal peptides | Ubiquinone and other terpenoid-quinone biosynthesis	5	4	0.47	5.34E-07	0.34	0.50
10	Bacterial secretion system	4	4	0.45	5.31E-05	0.64	1.00
11	Two-component system	90	3	0.37	0.00E+00	0.33	0.02

12	Glycerophospholipid metabolism	6	3	0.40	6.79E-13	0.42	0.11
13	Lipopolysaccharide biosynthesis	4	3	0.23	2.64E-10	0.43	1.00
14	Arginine and proline metabolism	6	2	0.27	3.88E-21	0.40	1.00
15	ABC transporters	5	2	0.13	1.84E-21	0.43	1.00
16	Purine metabolism	8	2	0.38	5.53E-25	0.33	0.25
17	Two-component system	10	2	0.23	2.67E-68	0.31	0.25
18	ABC transporters	4	1	0.00	1.77E-18	0.30	0.50
19	Two-component system	17	1	0.00	0.00E+00	0.27	0.25
20	Methane metabolism	6	1	0.00	1.03E-43	NaN*	1.00
21	Two-component system	8	1	0.00	5.55E-80	0.20	0.14
22	Protein processing in endoplasmic reticulum	4	1	0.00	1.77E-18	NaN*	0.02
23	Phenylalanine metabolism	10	1	0.00	2.85E-127	0.15	0.04
24	Methane metabolism	4	1	0.00	1.77E-18	NaN*	1.00
25	Arginine and proline metabolism	4	1	0.00	1.77E-18	0.07	0.09
26	ABC transporters	4	1	0.00	1.77E-18	0.06	0.25
27	Two-component system	10	1	0.00	2.85E-127	0.06	0.14
28	Cysteine and methionine metabolism	4	1	0.00	1.77E-18	0.16	0.33
29	ABC transporters	6	1	0.00	1.03E-43	0.20	0.33

*NaN values for pathway coherency indicate that KOs in this module were not represented in the set of microbial reference genomes examined.

3.4.2 Assembly modules vary in phylogenetic and pathway coherency

Many well-documented metabolic pathways in literature have been defined based on the functions present in a single organism. However, since the modules that we detect are defined based on co-occurrence across *metagenomes*, rather than individual species' genomes, they thus may comprise KOs from a variety of taxa, and may differ significantly from currently defined KEGG pathways. Therefore, we assessed module structure according to two criteria – phylogenetic coherency and KEGG pathway coherency.

Phylogenetic coherency of a module was defined as the mean similarity of constituent KOs in terms of their presence or absence across known microbial species. Assessing coherency via a panel of >2000 microbial reference genomes (see 3.3 Methods), We find that while a number of modules indeed show increased probability of fully contained

within individual microbial genomes, some modules contain KOs that have fairly low phylogenetic coherency (Fig. 3.2) – these KOs are no more likely to come from the same microbial genome than a background distribution of all pairs of functionally related KOs. This suggests that the presence of these modules in a given microbiome may rely on a number of co-occurring microbial *species*.

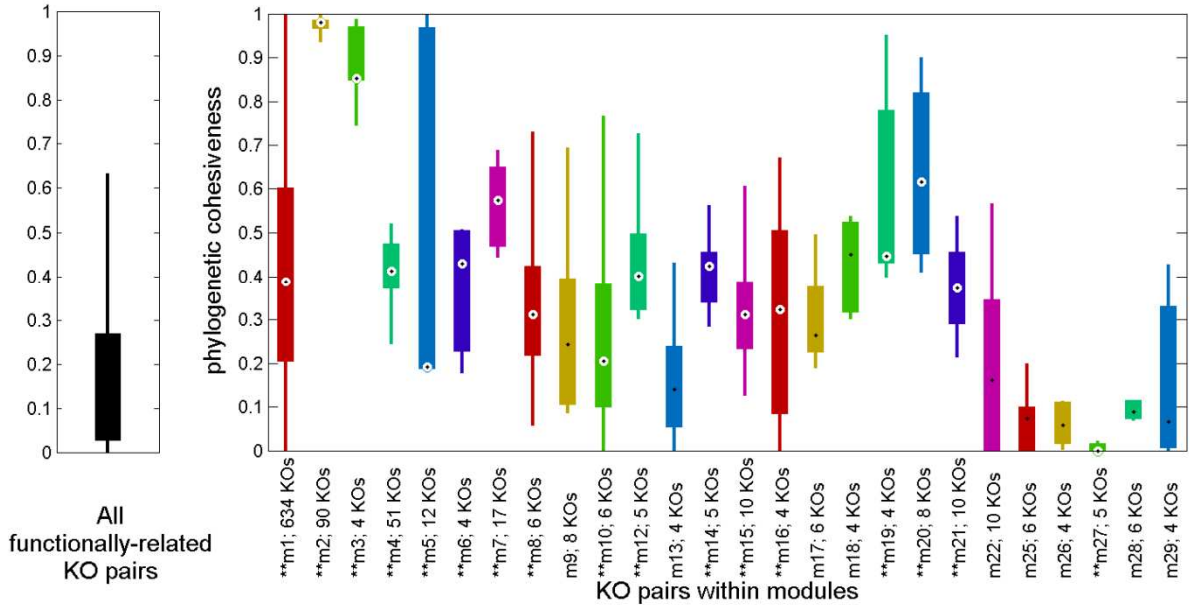


Fig. 3.2 Phylogenetic coherency of detected modules. Phylogenetic coherency (defined as the mean of pairwise similarity scores of KO presence/absence profiles across a large set of microbial reference genomes) was calculated for each module. The distribution of scores within each module is shown as a colored boxplot, and compared to the distribution of scores for all functionally-related KO pairs. Modules whose scores are significantly different from this background distribution are marked with **.

KEGG pathway coherency was assessed by first creating pie charts representing the relative representation of different KEGG pathways across the KOs in a module. The coherency of a module was defined as the mean fractional area of all slices in its pie chart. This indicates the degree to which a detected module overlaps with the boundaries of KEGG pathways. Low pathway consistency indicates that a set of KOs from multiple

currently defined pathways co-occur as a functional unit, and may represent a new *metagenomic* pathway. We find a number of modules with low KEGG pathway coherency. For example, module 23 has a coherency of just 0.04, and is composed of a set of 10 KOs with diverse pathway associations, including pathways related to amino acid, lipid, and carbohydrate metabolism (Fig. 3.3). Other modules have high pathway coherency, such as module 20, which consists of a set of 8 KOs all contained within KEGG's methane metabolism pathway. Notably however, these KOs are just a subset of the defined pathway, and may represent one specific aspect of methane metabolism, or one of many alternative paths.

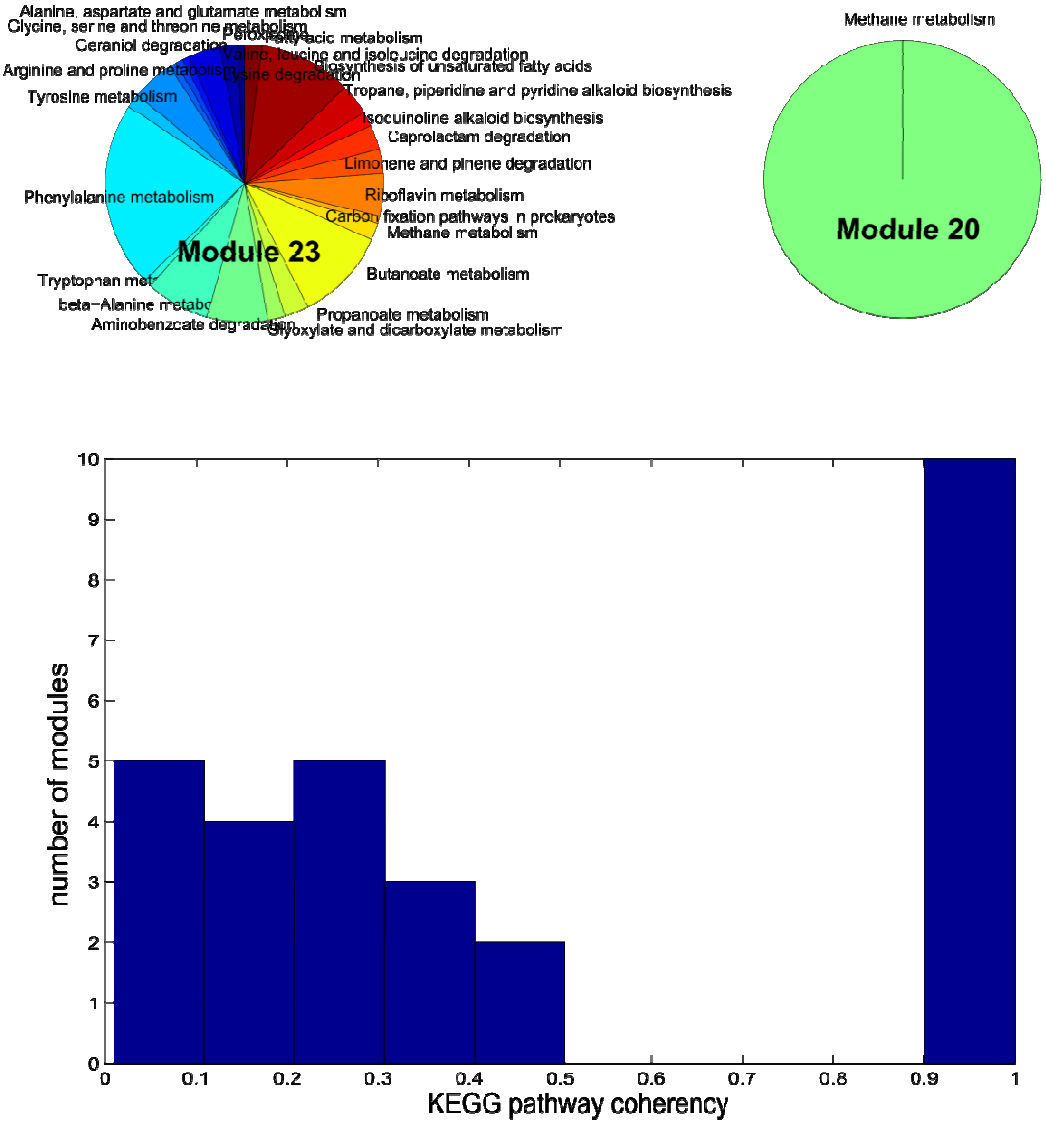


Fig. 3.3 Pathway coherency of detected modules. KEGG pathway coherency (defined as the mean of KEGG pathway relative representation scores within a module), were calculated for each module. A histogram shows the distribution of pathway coherency scores, with the pathway relative representation of two example modules displayed above as pie charts.

3.4.3 Assembly module representation varies across mammalian microbiomes

We find in general that each mammalian microbiome is composed of a unique set of assembly modules. Certain modules, however, appear to be essentially universal across

the mammals in our dataset. In fact, the largest module by far was a set of 634 KOs of which no less than 76% were present in any of the 39 samples. The KOs comprising this module came from a diverse set of pathways, including a number of housekeeping genes, core metabolism genes, and genes related to transcription, and translation. This large universal module reflects the known functional core already demonstrated in previous studies. 2 other smaller modules, one composed of 51 ribosomal KOs, and another containing 5 riboflavin metabolism KOs were fairly universal as well (represented in >35/39 samples).

Outside of this core, the majority of modules were found in only a relatively small subset of samples (Fig. 3.2). Among modules shared by more than one sample, most were found primarily in omnivores and carnivores, and few modules were shared among herbivores. These shared modules comprise a number of different carbohydrate and amino acid metabolism KOs (Table B.1). For example, modules 4,5 and 6 contain sets of KOs related to the transport of diverse starches including mannose, sucrose, beta-glucoside, and fructose (module 4), glucose, trehalose (module 5), and lactose, and cellobiose (module 6). Interestingly, none of the modules appear to be specific to herbivores. Representation of these three modules across omnivores and herbivores is high overall, yet certain individual samples lack one or two of these modules, potentially representing differences in dietary starch preferences. Importantly though, all omnivores and carnivores (except the two Baboon samples) contain at least one of these starch transport modules, while the majority of herbivores do not contain any. Modules 7 and 8, however, appear to be essentially exclusive to omnivores (but are also present in the echidna, a phylogenetically distinct anteater-like mammal, whose diet consists of insects). Module 7 consists of KOs involved in the final steps of peptidoglycan biosynthesis, while module 8 consists of four signal transduction KOs involved in the response regulation of OmpR and NarL. The presence of other modules appears to be due to an interplay of host diet and phylogeny. For example,

modules 10-15 are only found in two specific clades, consistently present in lion, spectacled bear, and black bear samples. Notably, the two lion samples in the dataset share a similar pattern of module presence, despite the fact that these two samples were sequenced with different technologies (GS-FLX and GS Titanium, respectively). Similar validation can be seen in the two chimp samples and the two okapi samples. Finally, modules 18-29 are each unique to a single sample. Many of the KOs in these modules are related to energy metabolism and signal transduction, while few are associated with carbohydrate or amino acid metabolism. Most of these sample-specific modules may equally represent variation between species, or variation relevant to only the sequenced representative of the species. However, the two samples from hyrax microbiomes, one from the San Diego zoo, and one from the St. Louis zoo, allow greater resolution. Module 20 is only found in the hyrax from the St Louis zoo, and contains a set of 6 methane metabolism KOs. While the presence of this module could be attributed to any number of environmental factors, such as differences in zoo feeding, climate, or housing facilities, this indicates that our method is capable of detecting both species-specific and organismal functional variation.

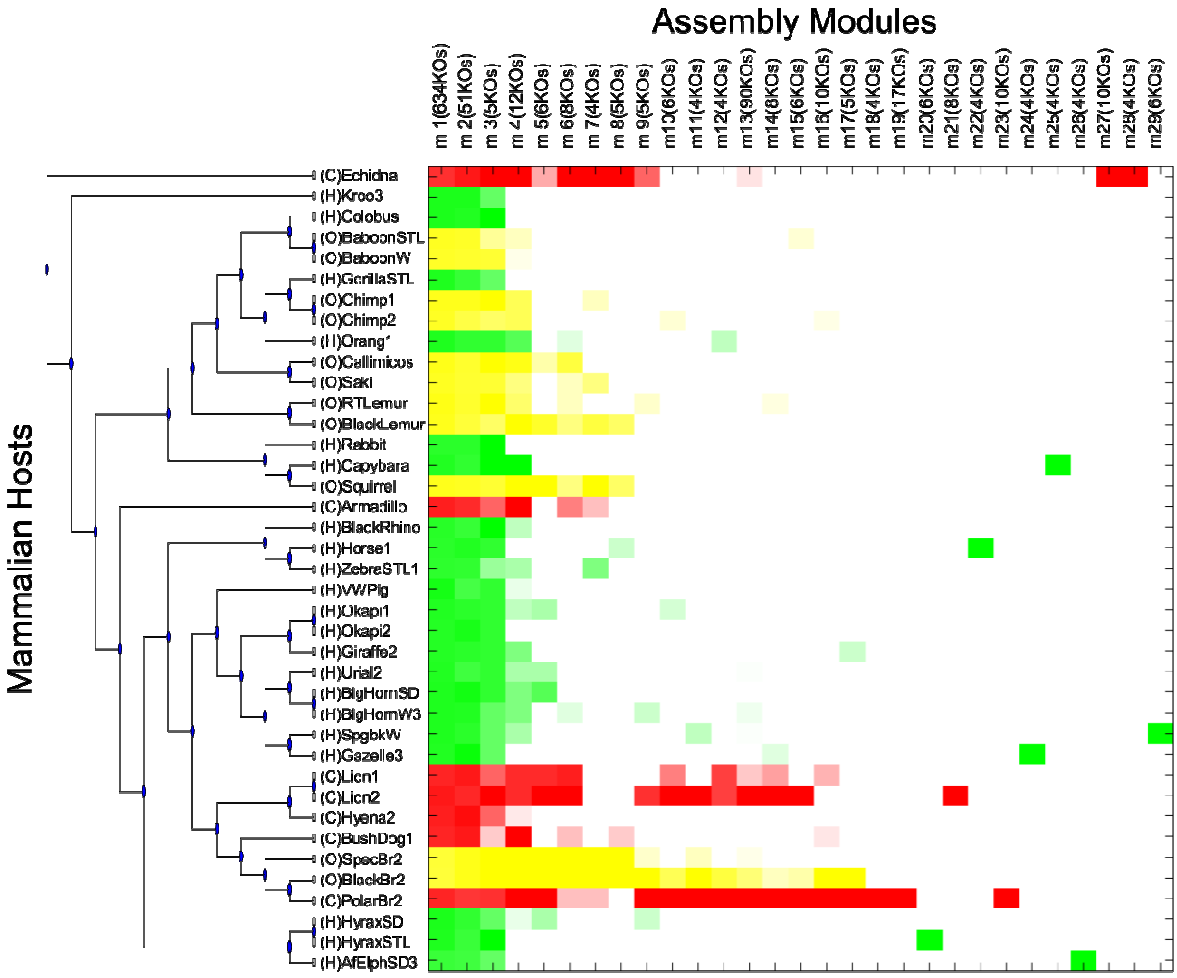


Fig. 3.4 Map of metagenomic assembly module representation across 39 mammalian hosts. Each square represents the representation of a give module (y-axis) in a given host (x-axis). Squares are colored by host diet (red:carnivore; yellow:omnivore; green:herbivore), and shaded according to the percent of KOs in the module that were detected in the sample (white:0%, dark:100%) . Mammalian hosts are ordered by their phylogeny according to the NCBI Taxonomy database. Modules are ordered by their prevalence across samples. See Table 3.1 for module descriptors.

3.5 Discussion

The mammalian lineage has evolved over millions of years, yielding a diverse set of species with various physiological, dietary, and geographical niches. This variation is reflected in the mammalian genome, where identification of co-evolving sets of genes has proved illuminating for understanding functional adaptation[106]. However, the impact of

mammalian evolution on the mammalian microbiome has only just begun to be studied. Current research indicates that both species and gene composition of the mammalian microbiome have been shaped primarily by host diet. However, description has thus far been limited to variation of overall gene profiles or individual genes; co-evolution within the mammalian *metagenome* has not yet been examined. Here we have undertaken a first pass at defining the elements of mammalian metagenome evolution, termed *metagenomic assembly modules,* by identifying sets of functionally-related co-occurring microbial genes across a diverse panel of mammalian microbiomes.

We have identified 29 gene sets that meet our criteria for assembly modules, collectively encompassing 944 individual genes. First and foremost, the existence of these modules, and our ability to detect them even using a simple search algorithm, indicate that higher-order structure and interactions may indeed contribute to patterns of microbiome evolution. At broad evolutionary scales, the gain or loss of metabolic genes is not an independent process – rather functional relationships between genes modulate their presence in a given microbiome.

The strong co-occurrence patterns of the genes in the modules we detect could be explained by a number of factors. We expect, for example, that genes in the same microbial genome would have strong co-occurrence signals. However, assessing the phylogenetic coherency of the modules, we find that many modules contain genes no more likely to be from the same microbial genome than any randomly chosen pair of functionally related genes. Thus, species interactions may play a part in the co-occurrence patterns we see. Furthermore, many of the modules contain KOs that are represented in a diverse set of KEGG pathways, suggesting that current pathway definitions may not be entirely appropriate for assessing the functional components of metagenomes.

Furthermore, we find that modules exhibit intriguing variation in representation across mammalian microbiomes. Three core modules appear to be universal, but many of the remaining modules are represented much more frequently in carnivores and omnivores than herbivores. By contrast, module representation in herbivores tends to be much sparser and more species-specific. This dichotomy could represent differences in the diversity of resources consumed by different diet groups – whereas meat-eating may require a similar set of metabolic capabilities regardless of the meat source, the broad range of plants and grains consumed by herbivores may give way to more defined metabolic functional niches. These results are especially interesting in light of recent discoveries indicating that the common ancestor of all terrestrial herbivores was a carnivore, and that a plant-based lifestyle arose only after subsequent adaptation[107], as it seems that herbivore microbiomes not retained many of the modules that are widespread across their carnivorous relatives.

The modules that we have identified represent the most cohesive gene groups in terms of co-occurrence. However, many more modules may exist that could be detected using a larger panel of samples, more sophisticated detection schemes, or more relaxed parameters. Future work in this area may also entail investigation of microbiome evolution at smaller time-scales. A number of recent studies examining shifts in microbiome gene content within a single host over time in response to changes in diet [84], [108], or within a single divergent host species [109], [110] have revealed functional coherence in the classes of genes that change the most. Such results suggest that functional relationships between genes indeed play a formative role in metagenome evolution; it will be of interest to determine whether the same or similar modules can be detected at these scales. Furthermore, a more precise understanding of the way a microbiome can change to provide new functional capacities may be crucial for future efforts in microbiome manipulation. For

example, defining the metagenomic modules best suited for a specific diet may provide new ways to combat malnourishment and make the most of resources in times of scarcity, new ways to overcome certain food intolerances, or even allow humans to subsist on new energy sources. Continued study in this vein will hopefully reveal even deeper insights about the evolution of microbiomes and how functional relationships shape metagenome composition.

4. Strain-level copy number variation is widespread across the human gut microbiome

This chapter is based on the following manuscript, which was accepted to *Cell,* and is currently in press:

Sharon Greenblum, Rogan Carr, Elhanan Borenstein. *Strain-level copy number variation is widespread across the human gut microbiome*.

4.1 Summary

Within each bacterial species, different strains may vary in the set of genes they encode or in the copy number of these genes. Yet, taxonomic characterization of the human microbiota is often limited to the species level or to previously sequenced strains, and accordingly, the prevalence of intra-species variation, its functional role, and its relation to host health remain unclear. Here we present a first comprehensive large-scale analysis of intra-species copy number variation in the gut microbiome, introducing a rigorous computational pipeline for detecting such variation directly from shotgun metagenomic data. We uncover a large set of variable genes in numerous species and demonstrate that this variation has significant functional and clinically-relevant implications. We additionally infer intra-species compositional profiles, identifying population structure shifts and the presence of yet uncharacterized variants. Our results highlight the complex relationship between microbiome composition and functional capacity, linking metagenome-level compositional shifts to strain-level variation.

4.2 Background

The human gut microbiome plays an important role in host metabolism, immunity, and drug response, and has a tremendous impact on our health [111]–[113]. Numerous comparative studies, aiming to characterize the contribution of the microbiome to human health have already demonstrated marked shifts in the relative abundance of various species, genera, or phyla in

various disease states [43], [114]–[116]. Clearly, however, each microbial species represents many different strains which may encode considerably different sets of genes and a different number of copies of each gene (reflecting for example, gene deletions and duplication events). Such intra-species variation endows each strain with potentially distinct functional capacities. Studies of individual isolates of cultured species have indicated, for example, that strains often differ in virulence [117]–[119], motility [120], nutrient utilization [121], and drug resistance [118]. Accordingly, the true functional potential of a microbiome cannot be inferred from species composition alone. Recent efforts to catalog the relative abundance of known strains in human microbiome samples [122] may recover some of these differences but are limited to sequenced reference genomes and are not be able to identify novel, yet to be sequenced, variation. Gene-centric shotgun metagenomic studies on the other hand, may identify genes or pathways that are differentially abundant across samples, but cannot necessarily attribute these shifts to specific species or strains. Specifically, it is often unclear how much of the observed variation in gene composition is due to variation in the abundances of species and how much is contributed by intra-species gene variation. Indeed, conflicting results have been reported, with trends identified among species profiles that are often poorly translated to gene profiles and vice versa [43], [123]. It is therefore not yet clear how prevalent gene-level intra-species variation is in the human gut, whether such variation is adaptive and affects specific functions, and how much of this variation has already been captured by reference genomes.

There is some evidence to suggest that variation among strains is common in the human gut. Several studies have focused more specifically on nucleotide-level strain variation, assessing for example, the prevalence and stability of single-nucleotide polymorphisms in 101 genome clusters across numerous metagenomes [124], or the level of sequence diversity across multiple near-complete genomes from two bacterial species variants obtained by single-cell sequencing [125]. Other studies have taken steps to associate sequence variation with

gene-level differences, identifying for example, areas of variable coverage and the coordinated loss of genes from specific gene families within the *Streptococcus mitis B6* genome [10], or a diverse array of strain-specific adhesion-like protein genes across 20 isolated and cultured strains of *Methanobrevibacter smithii* [126]. Additional studies have used extensive manual genomic reconstruction and assembly to track strain-resolved shifts in *Actinomycetaceae* in the relatively low-complexity premature infant gut microbiome over time [127], to detect differences related to antibiotic resistance, transport, and biofilm formation among three strains of *S. epidermis* [128], or to identify the variable presence of genes involved in transport, motility, carbohydrate metabolism, and virulence in two distinct strains of *Citrobacter* [129]. These gene-level variation studies, however, mostly report small-scale or anecdotal results, focusing on one or a small number of species and often on specific gene families. A high-throughput comprehensive analysis of gene-level variation, tracking the presence/absence and copy number of every gene across a large array of species in the human gut is therefore needed in order to more fully appreciate the extent and functional implications of strain variation in this complex microbiome.

To address this challenge, here we establish a rigorous and robust pipeline to estimate the copy number of each gene in a large set of prevalent gut microbial species directly from metagenomic shotgun data, and furthermore to detect copy number variation across samples. We carefully calibrate this pipeline to confirm that it can successfully estimate the copy number of individual genes in individual species on a large scale. Applying this pipeline to 109 metagenomic samples from a recent study of the gut microbiomes of lean, obese, and inflammatory bowel disease (IBD) afflicted individuals, we estimate the copy number of more than 4,000 gene groups across 70 species in each of these samples, and demonstrate the presence of widespread copy number variation within many gene-species pairs. We find that specific functions are especially prone to copy number variation, including functions relevant to

a community lifestyle and adaptation to the gut environment, and further detect associations between strain variation and host phenotype. Finally, we demonstrate that these copy number estimates can be used both to model the composition of known strains within each sample, and to offer insight into complex population structures suggesting the presence of yet uncharacterized species variants.

4.3 Methods

4.3.1 Metagenomic samples

Metagenomic data were obtained from [130], a study characterizing the gut microbiome of Danish and Spanish individuals, including individuals afflicted with obesity or IBD. Illumina-derived shotgun reads (75bp) from 109 samples were downloaded from the DDBJ ftp site [131] at ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/ERA000/ERA000116/. Additional samples from this study sequenced with 44bp reads were not included in our analysis.

4.3.2 Reference genomes and annotation

A list of 261 dominant and prevalent human gut microbial strains, grouped into 101 genome clusters based on sequence similarity of 40 marker genes, was obtained from [124]. Reference genomes with corresponding taxon IDs were downloaded from NCBI's Genbank when present, or from NCBI's draft genome submissions. One genome was not present in either and was omitted from further analysis (*Salmonella enterica subsp. enterica serovar Paratyphi B*, taxonID: 272994). Nucleotide contig sequences, gene calls, and amino acid protein sequences were downloaded for each genome, and protein sequences were annotated with KEGG orthologous groups (KOs) using BLASTp against KEGG v. 8/6/2012 limited to prokaryote peptide sequences. Proteins with multiple best hits were annotated with all best hits, weighted by the number of hits for each KO.

4.3.3 Alignment of reads to reference sequences and calculation of KC coverage

Shotgun short reads from the 109 metagenomic samples were aligned to the 260 reference genomes using BWA. Extensive simulations were performed to determine appropriate mapping parameters and identify reliable mapping targets (Fig. C.1, C.2; Appendix C). Each read was mapped to the reference sequence(s) with the smallest edit distance, weighted by the number of tied hits. Reads with a minimum edit distance > 5, or reads mapping equally to more than 75 regions were considered unmapped. In total, 2,469,102,286 reads were mapped to one or more reference genomes with these parameters. Average coverage over each gene region was determined using samtools [132] and the coverage of each KC (KO-cluster pair) was obtained by summing over all genes annotated with the same KO and genome cluster.

4.3.4 Calculation of copy number estimates

To translate the coverage values described above to copy number estimates, the coverage of each KC in each sample was normalized by the cluster's average coverage over a set of 13 marker KOs (Fig. C.3B; Appendix C). The resulting values, V_{kcs}, represent the estimated copy number of each KO *k*, in each cluster *c*, and in each sample *s*. 'Detectable KCs' in a sample were defined as those with V_{kcs} >= 0.5. 'Detectable clusters' within each sample were defined as those with at least 12 detectable marker KCs and an average coverage across the 13 markers KCs >= 1. KCs that were not detectable in any sample were removed from the analysis.

4.3.5 Detection of highly variable and set-specific variable KCs

For each KC present in each cluster detectable in at least 10 samples, the median copy number across samples and the MAD (median absolute deviation) from this median were calculated. The MADs of all 40,088 KCs formed a skewed distribution with a mean

0.1716 and standard deviation 0.2315. KCs with a MAD more than 2 standard deviations from the mean (MAD>0.6346) were considered *highly variable*. KCs in which at least 10% of samples had a copy number that exceeded the median copy number by this threshold were considered *set-specific increased variable KCs*. *Set specific decreased KCs* were similarly identified as KCs in which at least 10% of samples had a copy number that was lower than the median copy number by this threshold.

4.3.6 Functional over-representation of variable KCs

The KEGG database was used to associate each KC with functional pathways, modules, and/or BRITE term annotations. Within each genome cluster, a hypergeometric enrichment test was used to assess the over-representation of either highly or set-specific variable KCs among KCs associated with each KEGG pathway, module, and/or BRITE term.

4.3.7 Detection of host state-associated KCs

Samples were labeled as obese, IBD, or healthy according to the data obtained from [130]. Samples that were labeled as both obese and IBD were omitted. A KC was defined as obesity-associated if the copy numbers in obese samples were significantly higher or significantly lower than the copy numbers in non-obese samples, according to a two-sample ttest (FDR-corrected $p<0.05$). IBD-associated KCs were similarly defined.

4.3.8 Copy number profile deconvolution using least squares regression and principal coordinate analysis

For each sample, a non-negative least-squares linear regression analysis was performed to obtain the a linear combination of reference genomes in each multi-genome cluster optimally explaining the copy number estimates of set-specific variable KCs calculated for that cluster in that sample. The regression was constrained such that the sum of genome weights for each sample and cluster equaled one. Prediction error was defined by calculating the R^2 value for each sample. A principal coordinate analysis was also

performed for every genome cluster, operating on the pairwise Euclidian distance matrix of set-specific variable KC copy numbers in each sample and each sequenced reference genome.

4.4 Results

4.4.1 A pipeline for calculating genomic copy number estimates in metagenomic samples

We developed a novel pipeline to confidently detect variation in gene content and gene copy number in a large set of prevalent human gut microbes directly from metagenomic data (Fig. 4.1; and see 4.3 Methods). Briefly, this pipeline works as follows. Shotgun metagenomic short reads were first aligned to a set of reference genomes representing dominant and prevalent gut microbiome strains. To account for the potentially multiple genomes available for each species in this reference database, genomes were grouped into clusters using a previously introduced sequence similarity-based method [124]. We used extensive simulations to carefully select alignment parameters and confirmed that that with these parameters, reads mapped to the correct region and correct genome cluster, while reads from genome clusters not represented in our reference database remained unmapped (Fig. 4.2A; Fig. C.1; Appendix C). In parallel, gene coding regions from all reference genomes were aligned to the KEGG database and associated with one or more KEGG orthology groups (KOs). We further performed extensive analysis to identify reference genomes and KOs for which alignment-based mapping had low confidence and omitted such genomes and KOs from downstream analysis (Fig. C.2; Appendix C.1). For each sample, coverage across regions corresponding to the same KO in the same genome cluster were summed and the average coverage of 13 single copy marker genes (K03043, K02996, K02952, K02871, K06942, K02982, K02933, K01887, K02986, K02863, K01873, K02994, K02876; see Fig. S3B for a list of KO names), carefully selected for their universality, mapping accuracy, and coverage consistency (Fig. C.3) was used to convert

the calculated coverage of each KO in each cluster to a copy number estimate (Appendix C.1). Overall, this process estimated copy number V_{kcs} of each KO *k*, in each genome cluster *c*, detected in each sample *s* (Fig. 4.1). Notably, copy number estimates represent an average across the various genomes associated with each cluster that are present in the sample and across the potentially multiple genes associated with each KO. We further performed an analysis of an extensive synthetic dataset to confirm that this scheme accurately recovers species abundances and copy number values (Fig. C.4A-B; Appendix C.1).

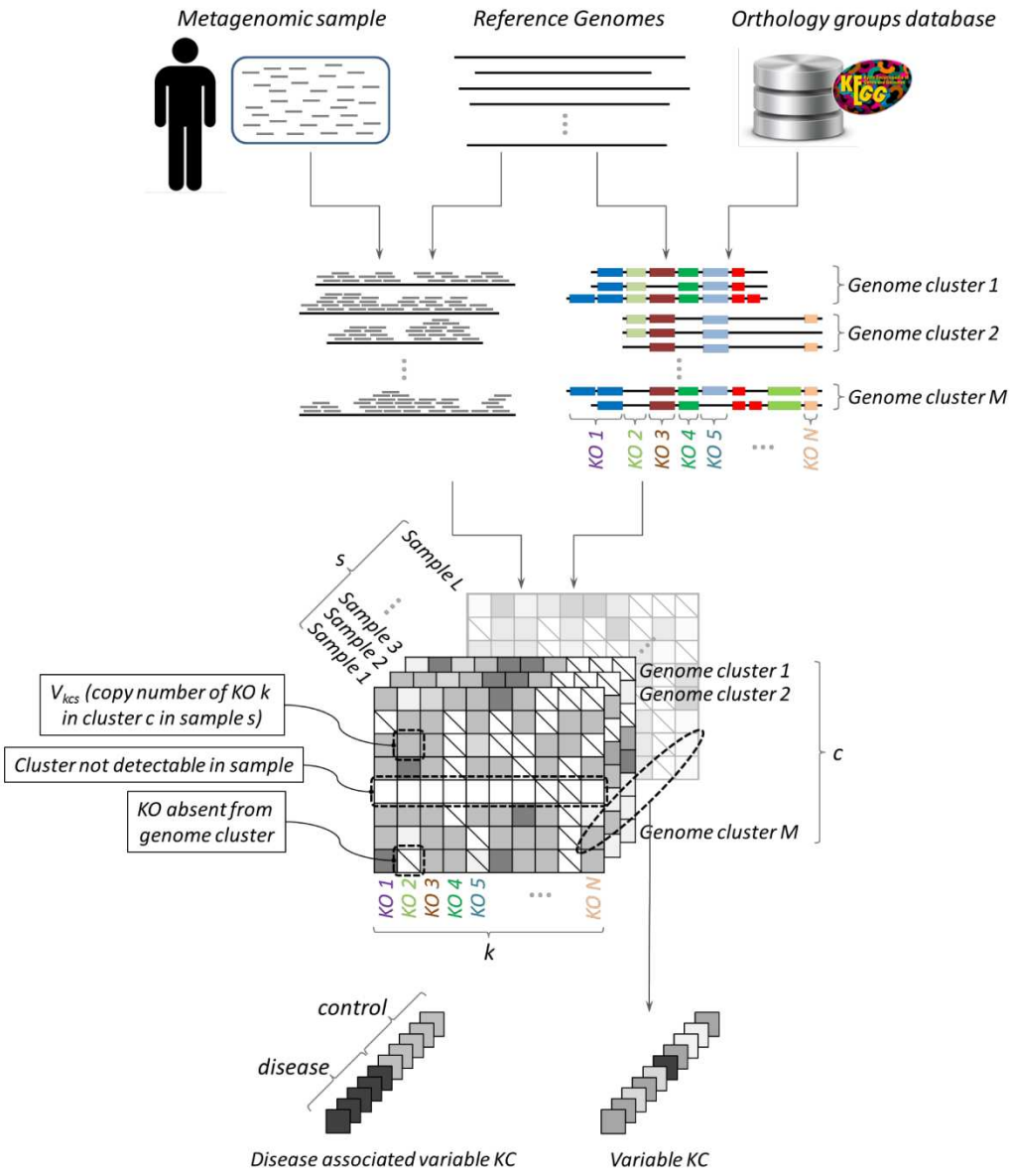


Figure 4.1: Schematic of analysis pipeline. Reads from metagenomic samples were mapped to KEGG- annotated reference genomes, grouped into species-level genome clusters. Comparing the coverage of each KO (KEGG orthology group), *k,* in each genome cluster, *c,* in each sample, *s,* to the average coverage of universal single copy marker genes was then used to calculate the copy number of each *k,c,s* triplet. A KC (specific KO in a specific genome cluster) was defined as variable if its copy number varied significantly across samples. Variable KCs whose copy number was associated with host state (obesity, IBD) were also detected.

We applied this pipeline to a dataset of 109 previously collected gut metagenomic samples [130], mapping in total >2.45 billion 75bp reads to 235 reference genomes grouped

into 96 genome clusters (Appendix C.1). The average coverage across the 13 marker genes (a proxy for the abundance of each cluster in each sample) varied considerably across clusters and between samples (Fig. 4.2*B*-*C*). To limit any downstream analysis to high confidence copy number estimates, we therefore considered only genome clusters with sufficient coverage in a sample (which we term 'detectable' clusters; Appendix C.1), identifying a total of 70 clusters that were detectable in at least one sample, with an average of 16 detectable clusters in each sample. Overall, this analysis was able to assign copy number values to ~1.5 million KO-cluster-sample triplets, estimating the copy number of thousands of KOs across a large array of genome clusters in more than 100 samples.

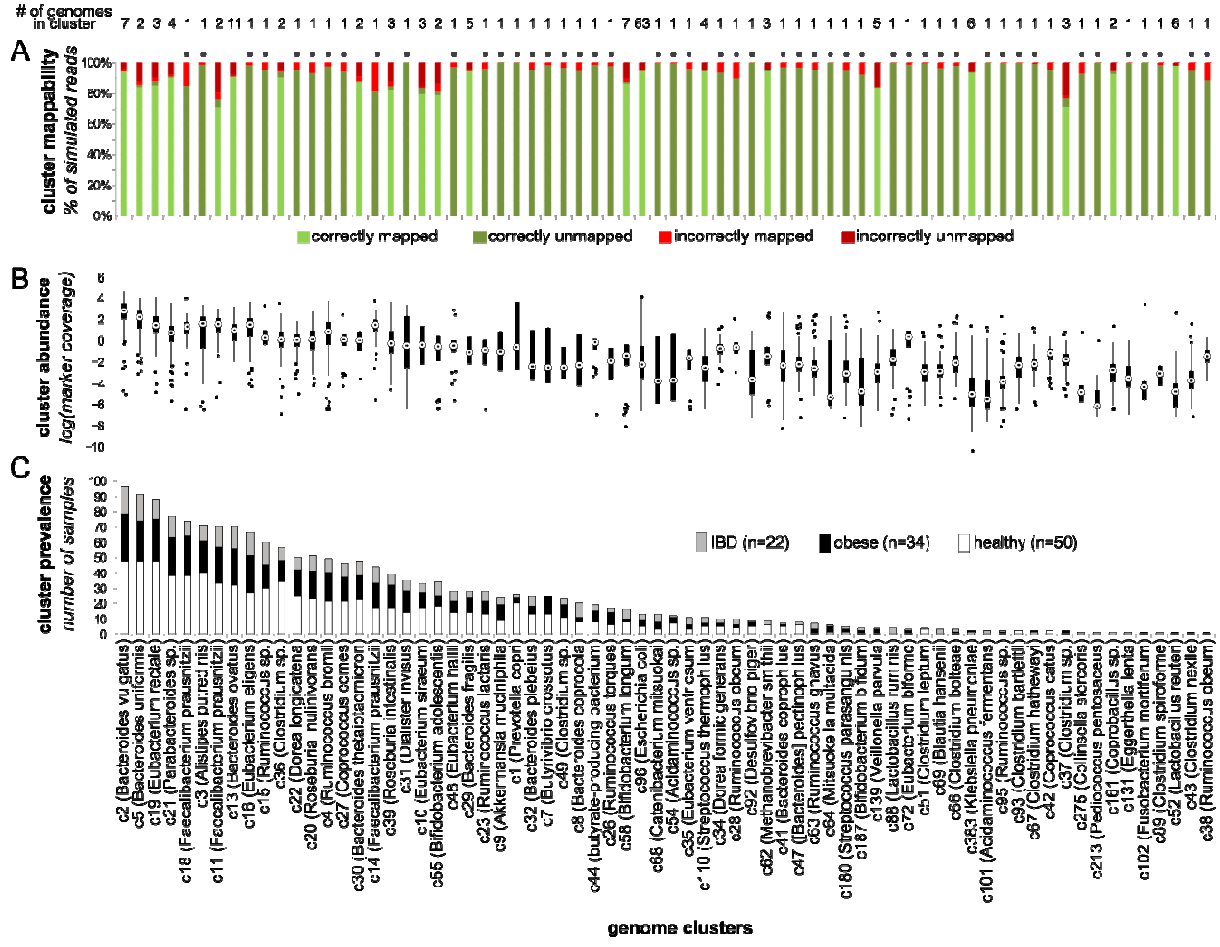


Figure 4.2: Cluster statistics. The mappability, abundance, and prevalence of each cluster across samples are shown in a 3 vertically aligned plots. (A) Cluster mappability, as determined by a large-scale simulation assay measuring the accuracy of mapping reads extracted from the cluster's genomes to a database in which the genome of origin was removed (see Appendix C.1 for details). Note that only reads from single-genome clusters (marked with a dot above the column) are expected to remain unmapped. (B) The distribution of each cluster's abundance across samples, as determined by the average coverage of 13 single-copy marker genes. (C) Cluster prevalence, defined as the number of samples in which the cluster was 'detectable' (see Appendix C.1), shown for each host group as a stacked bar plot.

This dataset of copy number estimates provides a first large-scale account of gene-level strain variation among organisms common to the human gut. Below, we mine this dataset to explore neutral and adaptive variation in this highly complex ecosystem in a manner that goes beyond species-level comparative analysis. Importantly, this dataset and the pipeline described above can serve as a valuable resource for future studies of compositional shifts in the human microbiome and in other environments, linking metagenome-level differences in gene abundance to genome-level variation.

4.4.2 Identifying genes with highly variable and with set-specific variable copy number

Given the copy number estimates obtained above, we set out to identify specific KOs in specific clusters (KO-cluster pairs, or KCs) whose copy number varied across samples. Notably, to detect variation, we compared the copy number of each KC across different samples rather than comparing the estimated copy number in any given sample to the copy number in a reference genome, avoiding spurious variation predictions that may result from annotation errors or bias in the set of reference genomes. Clearly, many KOs are encoded by only one or a few genome clusters and many genome clusters can be detected in only a few samples. To confidently detect copy number variation, we therefore only considered the 40 clusters that were detectable in at least 10 samples. We additionally excluded KCs with very low copy number values across all samples (see 4.3 Methods).

We first set out to identify KCs that exhibit extreme and prevalent variation across samples. Specifically, we calculated the level of inter-sample variation in the copy number of each KC, and defined as *highly variable* those KCs whose variation was at least two standard deviations greater than the average variation of all KCs (4.3 Methods). In total, this analysis detected 735 highly variable KCs spanning 261 KOs across 38 genome clusters (Fig. 4.3). The number of highly variable KCs in each cluster varied greatly, reaching up to 47 KCs in the *Roseburia intestinalis* cluster (representing 4.05% of the KCs in this cluster), with an average of 1.79% of the KCs in each cluster. We found no apparent relationship between the amount of variation observed in a cluster and the number of reference genomes in the cluster or the prevalence of the cluster across samples, but we did observe a tendency toward high variation in species from the *Firmicutes* phylum compared to other species (t-test p-value <0.05; see also Fig. 4.3). While the majority of highly variable KOs (57.1%) were variable in just one cluster, certain KOs were variable across many clusters, with some KOs variable in 10 or more clusters.

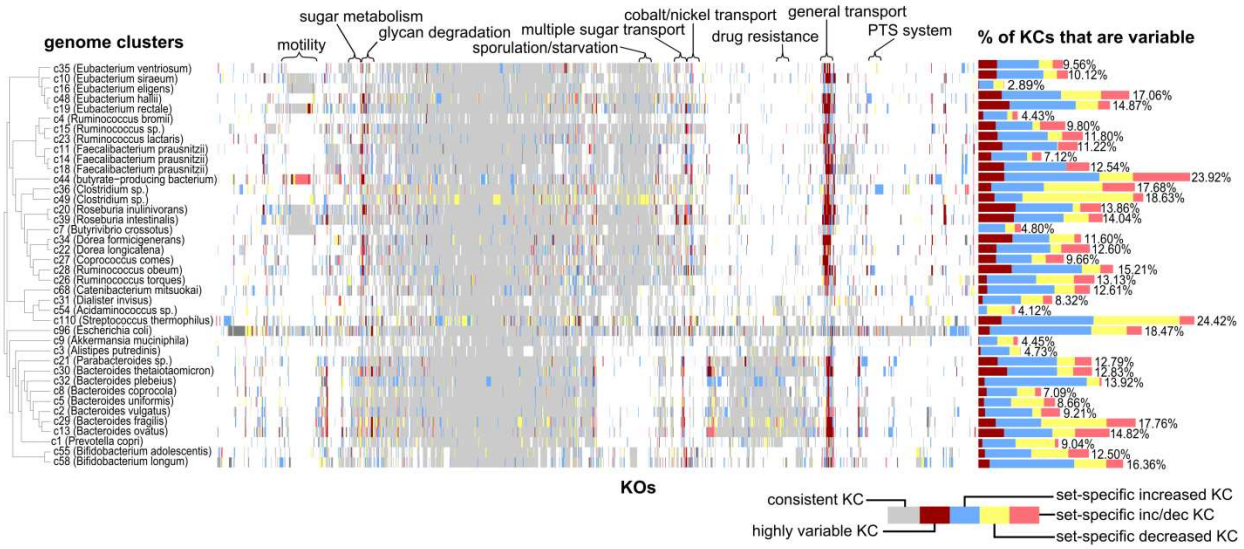


Figure 4.3: A map of variable KCs. A matrix map representing the status of each KO (x-axis; only KOs that were variable in at least one genome cluster are included) in each genome cluster (y-axis). Colored bars represent variable KCs (partitioned into variation types as defined in the text), while light gray bars indicate KCs with consistent copy number across samples, and KOs not present in a genome cluster are left white. Genome clusters are ordered by phylogeny and KOs are ordered by hierarchical clustering. The bar chart to the right of the map represents the fraction of KOs in each cluster identified as variable (again, partitioned into the different variation

types). Above the map, certain groups of functionally-related KOs are highlighted. The 314 KOs uniquely variable in the E. coli cluster were excluded due to space constraints.

The analysis above focused on KCs that exhibit extreme variation, and on KCs that vary greatly across *many* different samples. Variation within other genes, however, may be more subtle and may reflect, for example, adaptive variation that can be observed in only a small set of samples. We therefore set out to additionally identify *set-specific variable KCs,* wherein the copy number of a given KC was relatively constant across most samples but deviated significantly and consistently in a small subset of the samples (Methods). In this analysis, we further distinguished cases in which a KC exhibited a consistently high copy number in this subset of samples compared to all other samples (set-specific *increased* copy number) from cases in which a KC exhibited a consistently low copy number in this subset of samples (set-specific *decreased* copy number) or in which it exhibited increased copy number in one subset and decreased in another. As expected, we found that set-specific variable KCs were much more common than highly variable KCs. In total, our analysis detected 5,004 set-specific variable KCs covering 1,859 KOs across the 40 genome clusters examined (Fig. 4.3). In general, we observed more cases of set-specific increased copy number than of set-specific decreased copy number but this ratio shifted markedly across clusters, and in certain clusters (ie. *Clostridium sp.*, *Streptococcus thermophilus*) mostly set-specific decreased copy number KCs were observed.

4.4.3 Detected variation captures both known and novel strain variation

As validation of our pipeline and results, we compared the set of highly variable KCs obtained for each cluster to known variation among the cluster's sequenced reference genomes. Clearly, the reference genomes in our database do not capture the full extent of intra-species variation in the gut microbiome, and similarly our samples may not include all the variation present in our reference genomes. Yet, the set of detected highly variable genes,

which aims to include genes that vary frequently in their copy number across genomes, is likely to capture many instances of known variation in gene content among available reference genomes. Indeed, considering the 15 multiple-genome clusters in our database, a striking 81% of the detected highly variable KCs also vary in copy number across reference genomes (Fig. 4.4A). Moreover, in 7 of these clusters, *all* highly variable KCs also vary in copy number across reference genomes. Notably, 6 of these clusters contain at least 3 genomes, while the majority of the other clusters contain only 2, suggesting that more sequenced strains may be needed to fully capture the variation associated with these clusters (and with clusters for which only a single genome was available). Importantly, we demonstrated that a similar overlap can be observed when comparing predicted variation to known variation among genomes *not* included in our database, confirming that this overlap is not an artifact of the specific reference genomes used in our analysis (Fig. 4.4 B-C; Appendix C.1). Comparison of set-specific variable KCs to known variation across reference genomes again confirmed that the variation detected greatly overlapped with known variation observed across sequenced strains (Fig. C.5). Notably, however, set-specific variable KCs also included many instances of novel variation, suggesting that the set of reference genomes currently available does not capture the full extent of copy number variation in the gut.

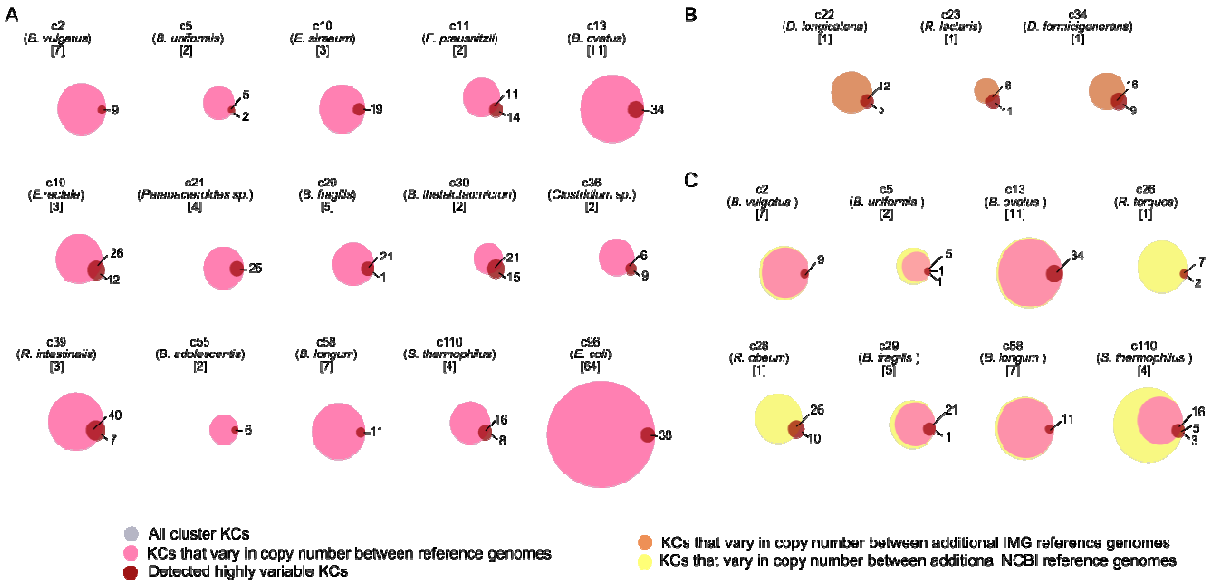


Figure 4.4: Comparison of highly variable KCs to known variation among reference genomes. In each Venn diagram, the gray circle represents the set of all KCs in a given genome cluster, the purple circle represents the fraction of those KCs exhibiting copy number variation across the cluster's reference genomes, and the red circle represents the set of KCs detected as highly variable. Overlap of the purple and red circles indicates correspondence between known and detected variation. In the inset, detected variation in three single-genome clusters is compared to known variation that was assessed by examining additional related genomes that were not used as a mapping target in our analysis (orange circle).

As additional validation, we compared our results to specific instances of known copy number variation detected across two manually-assembled genomes representing distinct strains of *Citrobacter* found in the deeply-sequenced gut microbiome of a premature infant [129], and observed a significant overlap in the sets of variable genes. Specifically, functional differences between the known strains included the presence or absence of fimbrial genes and genes involved in phenylacetate degradation. Within our dataset, we found that 13 of the 14 phenylacetate degradation KCs in the genome cluster containing Citrobacter genomes, and 7 of the 12 fimbrial KCs were identified as set-specific variable KCs. While it is not expected that our samples, obtained from European adults, would necessarily harbor the same strains found in a single premature infant, the similarities in the types of functions that are subject to increased or decreased copy numbers are intriguing.

4.4.4 Functions associated with variable genes

We examined whether the detected copy number variation was associated with specific functions in each genome cluster. We first used enrichment analysis to identify functions that were over-represented among the set of highly variable KCs in each cluster. We found that transport-related functions were overwhelmingly prone to high copy number variation (Table C.2). Specifically, 10 of the genome clusters analyzed were enriched for variation in KCs associated with transport annotations, including the general BRITE term ‘Transporter’, as well as more specific modules related to either sugar or iron complex transport. For example, within the *Bacteroides ovatus* cluster, seven of the cluster’s 66 transport-associated KCs were highly variable (Fig. 4.5), including all three KCs (K02013, K02015, K02016) involved in a specific iron complex transport system module (M00240). Interestingly, significant variation in sugar transport functions was only found among clusters in the phyla *Firmicutes* and *Actinobacteria*, while *Bacteroidetes* clusters were uniquely associated with variation in the iron complex transport system. Studies of cultured organisms from various environments and experimental evolution assays have suggested that loss, amplification, and acquisition of transport functions constitute a primary adaptive mechanism [133]–[136]; here we show that this flexibility in the copy number of transport genes likely extends to a considerable proportion of prevalent gut species, and that within this general class, specific transport genes may facilitate adaptation to the gut environment.

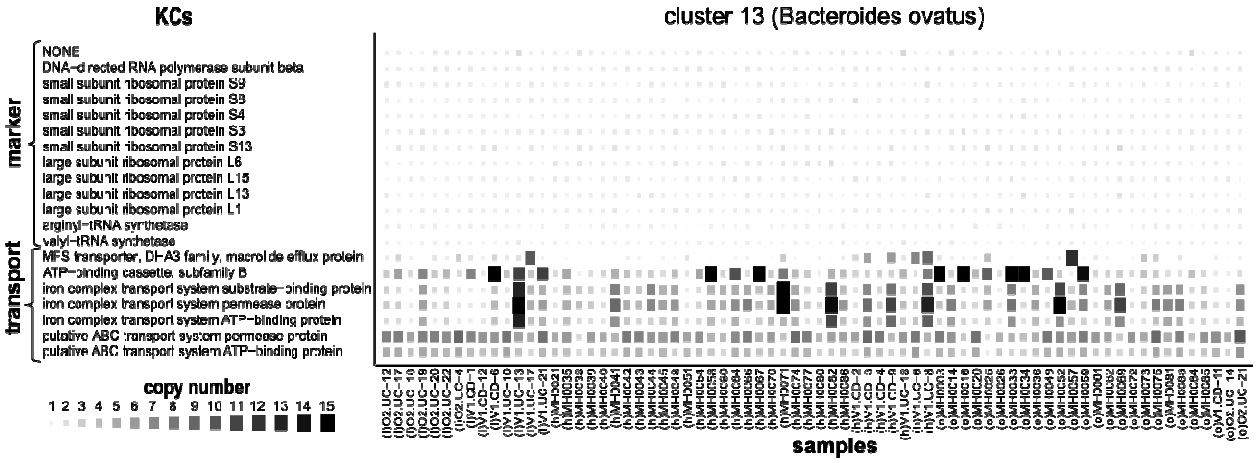


Figure 4.5: Copy number of highly variable transport KCs in *Bacteroides ovatus.* The size and color of each square represents the copy number of each highly variable KC within each sample. Samples are grouped by host state (I: IBD, h: healthy, o: obese). The copy number of the 13 marker KCs in this genome cluster and across the samples are also illustrated for comparison.

We additionally found that motility-related KCs were highly variable in the *Eubacterium rectale* genome cluster. Specifically, in this cluster 7 of the 38 highly variable KCs were bacterial motility proteins, of which 4 were structural flagellar components, 2 were involved in chemotaxis, and 1 was essential for twitching motility [137]. Motility proteins, and especially flagellar proteins, are widely associated with virulence and immunostimulation, and the gain or loss of flagellar components is believed to be an important adaptive mechanism for individual organisms [135], [138], [139]. Moreover, the variation in these 7 KCs was highly consistent within samples; most samples contained either detectable copies of all 7 KCs or no (or low number of) copies of all of these KCs (Fig. C.6). While we found no variation in the copy number of any of these genes among the three sequenced reference genomes included in the *Eubacterium rectale* cluster in our database, a recent study of 27 elderly gut metagenomes identified non-uniform coverage of genes involved in the flagellum biogenesis pathways of six *Eubacterium* and *Roseburia* species [140], suggesting that the current reference genomes may not capture the full dynamic range of these species.

Next, we considered the collection of set-specific variable KCs and examined their functional annotations. Interestingly, hierarchical clustering of set-specific variable KOs based solely on their variation profile across the 40 clusters revealed distinct groups of functionally-related genes that vary in a given genome cluster or within multiple clusters (Fig. 4.3). For example, a large set of genes related to cell growth and sporulation were all identified as set-specific variable KCs in the two genome clusters associated with *Clostridium sp.* Similarly, a set of sugar metabolism genes were all identified as set-specific variable KCs in *Roseburia intestinalis*, and a number of antibiotic resistance genes were identified as variable in multiple genome clusters, primarily those in the *Firmicutes* phylum. An enrichment analysis of functions associated with set-specific variable KCs in each cluster additionally revealed a number of important functions that are prone to copy number variation (Table C.2). For example, genes in the lipopolysaccharide biosynthesis pathway in *Dialister invisus* and *Clostridium sp.* were often observed with a higher copy number in a small set of samples (ie., set-specific increased copy number). Interestingly, variation within functions related to sugar metabolism (ie. KEGG pathways galactose metabolism, starch and sucrose metabolism, fructose and mannose metabolism, polyketide sugar unit biosynthesis) was observed primarily within *Bacteroidetes* clusters, while set-specific transport-related variation was almost absent from these clusters. Other functions enriched for set-specific variable KCs suggest transitions between virulent and neutral states, such as motility in *butyrate-producing bacteria, Eubacterium rectale*, and *Clostridium sp.*; streptomycin biosynthesis in *Acidaminococcus sp*.; lysosyme production in *Bacteroides ovatus*; the EHEC/EPEC pathogenicity signature in *Escherichia coli*; and secretion systems in *butyrate-producing bacteria*, *Clostridium sp*., and *Escherichia coli*. Within *Escherichia coli*, type II secretion system genes were identified as set-specific decreased copy number KCs, while type III secretion system genes were identified as set-specific increased copy number KCs. Overall, much of the observed variation appeared to be associated with the

way a species responds to and interacts with its surroundings, highlighting the strong adaptive potential of gut-associated bacteria.

Different cohorts could clearly harbor different sets of strains owing to an assortment of ecological or host-specific factors, and accordingly different sets of genes may vary in copy number in different datasets. Yet, the findings above suggest that genes associated with specific functions (and in specific species) may be more prone to copy number variation than others. We therefore wished to examine whether the set of KCs and functional classes detected in each species as variable are similar across different cohorts. To this end, we applied our mapping and analysis pipeline to a second dataset of 73 gut microbiome samples from a Chinese cohort (obtained from [52]), and compared the detected variation in this dataset to variation detected in our original Danish/Spanish cohort (Appendix C.1). Indeed, a marked overlap was observed both in the set of KCs identified as variable (65% for highly variable KCs and 75% for either highly or set-specific variable KCs) and in the set of functions enriched for copy number variation (59% for all functions and 68% for transport-related functions). These findings suggest that while the exact pattern of copy number variation may be personal and differ between different groups of individuals, certain genes and functions (e.g., those related to environmental adaptation) may be universally prone to variation.

4.4.5 Host state-associated variation

While much of the variation across strains may reflect neutral processes or transitory dynamics, some variation may represent adaptation to a specific host phenotype. To detect such potentially adaptive variation, we examined set-specific variable KCs and identified KCs in which the copy number in obese or IBD samples was significantly different than in healthy samples (4.3 Methods). In total, we found 24 KCs whose copy number was significantly associated with IBD and 3 KCs whose copy number was significantly associated with obesity (FDR<0.05; Table C.3).

Interestingly, a number of these KCs have been previously implicated in adverse host health states. For example, in our analysis, obesity was associated with a higher copy number of thioredoxin 1 (K03671) in both *Clostridium sp.* (Fig. 4.6A) and *Ruminococcus bromii.*, and indeed thioredoxin reductase was recently shown to be enriched in the cecal metaproteome of mice fed a high-fat diet [141]. Such results are consistent with thioredoxin's regulatory role in maintaining redox equilibrium and the demonstrated links between a high-fat diet and oxidative stress in mammals [142]. Additionally, in our analysis, the loss of a ubiquinone-reducing gene (K00349; *nqrD*) from *Bacteroides plebeius* was associated with obesity. A recent study in mice showed that supplemental ubiquinone reduced inflammation and metabolic stress accompanying a high-fat high-fructose diet by reducing the expression of certain genes associated with stress-response [143], while mice not receiving the supplement gained more weight than their counterparts. Importantly, however, ubiquinol, the reduced form of ubiquinone, has recently been shown to be the more readily absorbed and more active form of the compound [144], raising the possibility that loss of microbial ubiquinone-reducing capabilities from certain species may hinder the effectiveness and protective capacity of ubiquinone in the host. Other findings shed new light on previous studies of the disease-associated role of individual species, with variation associated with common disease hallmarks, such as pathogenicity-related secretion and antibiotic resistance. For example, in two clusters (*Faecalbacterium prausnitzii*, *Roseburia inulinivorans*) (Fig. 4.6*B*), increased copy number of a gene (K08217) coding for a major drug efflux protein known to play a role in antibiotic resistance was highly associated with IBD-afflicted individuals. Interestingly, these two butyrate-producing species have been associated with protective effects against IBD in past studies [145], [146]. The increased prevalence of toxin efflux pumps among the surviving strains detected here in IBD-afflicted individuals therefore may reflect a response to inter-species warfare, or a toxic gut environment in which non-resistant strains were inviable and populations decreased. In two

other clusters positively linked to IBD incidence [147] however, (*Bacteroides ovatus*, *Bacteroides uniformis*), increased copy number of a second multi-drug efflux pump (K03296) was also associated with IBD, potentially representing increased resistance and pathogenicity. Interestingly, a toxin secretion protein was also increased copy number in IBD samples in *Bacteroides uniformis*, further suggesting competition and warfare dynamics. See Table C.3 for a full list of disease-associated KCs. Further experimental analysis will clearly be required to determine exact mechanisms and distinguish cause from effect. Interestingly though, none of the obesity- associated KCs and only 3 of IBD-associated KCs were found to vary significantly in the Chinese cohort described above (among whom only one individual was obese, and none were reported as having IBD).

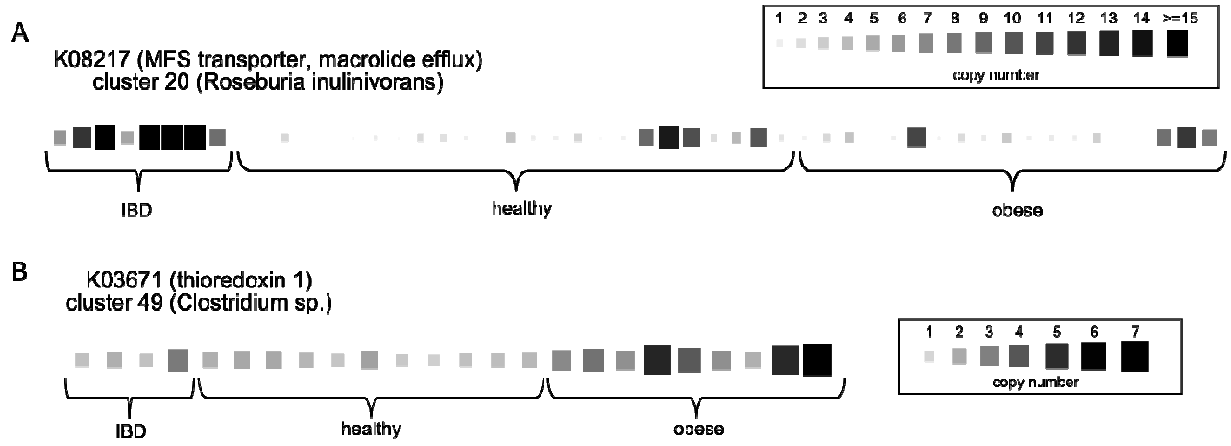


Figure 4.6: Copy number variation of host state-associated KCs. Two KCs whose copy number was significantly increased in samples from a specific host state are shown. The size and color of each square represent the copy number of the KC within each sample. (A) The copy number of an MFS transporter gene (K08217) in the *Roseburia inulinivorans* genome cluster is significantly increased in IBD samples. (B) The copy number of thioredoxin 1 (K03671) in *Clostridium sp.* is significantly increased in obese samples.

4.4.6 Deconvolution of microbiome composition and intra-species population structure

Clearly, the microbiome of different individuals can house multiple strains of the same species with potentially different relative abundances. Our copy number estimates for each cluster accordingly represent average copy numbers across the different strains in the sample. To date, however, most studies have characterized the species-level composition of the microbiome and relatively few large-scale studies have focused on variation in the composition of strains within each species. Here, we therefore examined whether our copy number dataset can be used to obtain insights into strain-level population structure. Notably, in this analysis we focused on the composition of strains within each genome cluster rather than on the abundance of the cluster itself within the sample.

First, we explored how well the copy number profiles obtained for each genome cluster in each sample can be explained by a linear combination of known reference strains, using a regression analysis to deconvolve these copy number profiles into a combination of the known strains included in our database (Methods). Obviously, these strains may not encompass the full set of strains present in the samples analyzed, yet, such analysis may be useful in examining what portion of the observed variation can be accounted for by known strains and what portion represents potentially novel variation. Indeed, examining the 15 genome clusters for which multiple sequenced strains were available, several interesting patterns were observed. Specifically, in well-characterized clusters with many sequenced genomes, the copy number profiles of most samples could be successfully explained by a linear combination of known strains. For example, in the *Escherichia coli* cluster which comprised 63 sequenced genomes in our database, 76% of the variation in copy number could be explained on average by these genomes (R^2=0.76±0.12). In this cluster, the inferred representation of each strain differed widely across samples, with some strains (ie. *Escherichia coli O111:H- str. 11128*) highly represented across multiple samples, and others were found in just one sample. However, in less well-characterized clusters with only a few known strains in our database, know strains

could explain in some cases just a small portion of the copy number variation observed in specific samples. For example, the four known strains of *Streptococcus thermophilus* could be used to explain a majority of the variation observed in most of the samples containing this genome cluster (R^2>0.5), yet failed to explain the variation observed in 4 of the samples (R^2<0), suggesting potentially novel variation (Fig. 4.7A).

To further compare copy number variation profiles and to examine such novel variation that may not be captured by known strains (including notably, in clusters comprising only one known strain), we used a principal coordinate analysis. This analysis revealed a complex and highly variable population structure within each cluster, with marked variation among samples indicating the prevalence of personalized variation. For a number of genome clusters, however, samples appear to group into distinct sets, potentially reflecting individuals with similar intra-species population structures (Fig. 4.7B). Moreover, by including the reference genomes in this principal coordinate analysis, we were able to distinguish variation observed across samples into previously captured vs. novel, yet-to-be-sequenced variation. For example, the principal coordinate plot for the *Streptococcus thermophilus* genome cluster (Fig. 4.7B) clearly demonstrates that while the copy number profiles of most samples clustered tightly with several known reference genomes, the 4 poorly-explained samples mentioned above clustered together and contained variation that was distinct from any reference genome. Such a pattern may indicate the presence of novel shared strains, providing a promising basis for targeted sequencing. Similar patterns were also observed in other clusters, where a distinct, tightly clustered subset of samples or individual samples exhibit markedly different copy number profile from that of any sequenced genome (Fig. C.7A-B). Overall though, each genome cluster exhibited a unique population structure across individuals, highlighting the complex suite of forces governing taxonomic composition in the gut [47].

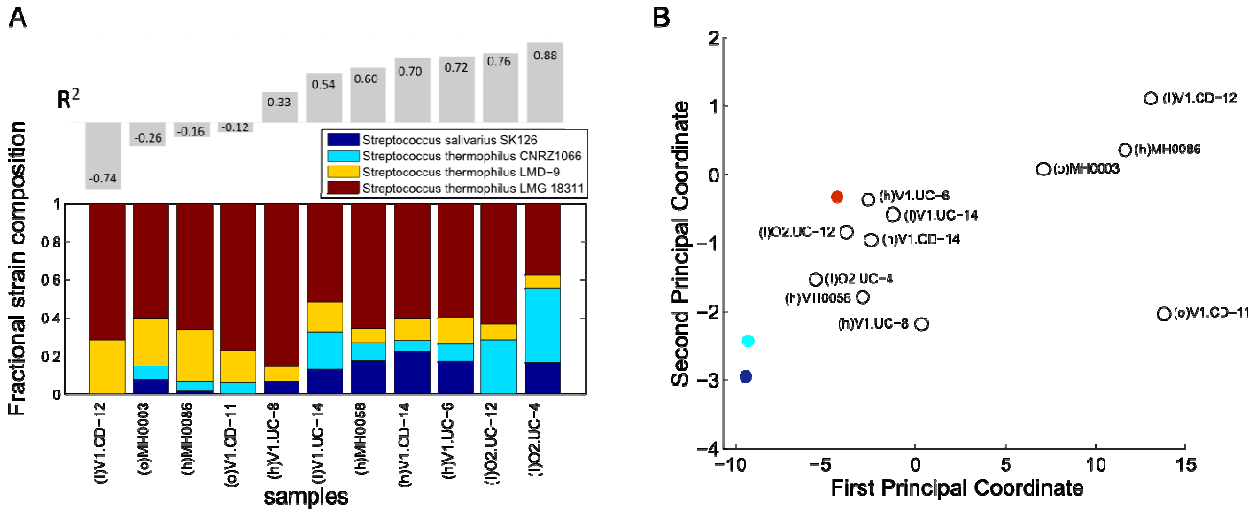


Figure 4.7: Predicted strain-level population structure within *Clostridium sp*. (A) A linear regression analysis was used to model the copy number profile obtained for cluster 110 (*Clostridium sp*) in each sample as a combination of known reference genomes, with prediction weights shown as stacked colored bars. Prediction accuracy (using R^2 values) is indicated above each bar. Samples with low or negative R^2 values potentially contain novel variation that cannot be explained by any combination of known reference genomes. (B) A principal coordinate analysis depicting the differences between the copy number profiles obtained for this genome cluster in the various samples (open circles), as well as the copy number profiles of reference genomes (filled circles).

4.5 Discussion

By and large, closely related organisms tend to encode similar sets of genes. This consistency is in fact often used to infer functional capacity from taxonomy [148], [149]. Clearly, however, this relationship between phylogeny and gene content is imperfect and each species represents a large collection of strains that differ in the set of genes they encode, the copy number of these genes, and ultimately, their functional capacity. Above we have focused on identifying instances where this relationship between microbial species and genes breaks, presenting a large-scale analysis of copy number variation in a diverse array of gut species. Our analysis has demonstrated that copy number variation is prevalent in the gut environment, with some species exhibiting significant copy number variation in >20% of their genes. Such variation may induce significant microbiome-wide shifts and may account for at least some of the observed discrepancies between trends observed at the species

levels versus trends measured at the gene level. Moreover, intra-species variation was shown to be especially prevalent in genes involved in specific functions, most notably functions that impact the way an organism interacts with its environment such as transport and signaling processes. This may suggest an adaptive dynamic by which certain species respond to changes in community composition or the gut niche and the potentially crucial role of the gut environment in shaping bacterial evolution [47], [150]. Other highly variable functions, such as lipopolysaccharide biosynthesis, cell motility, and secretion systems may represent changes in virulence as organisms respond to host immune responses. Interestingly, many of these same functions were highlighted in a previous study as those that were more difficult to accurately correlate with 16s data [148]. Our analysis further identified variable functions that may correlate with host states, exhibiting differential copy number in specific genomes. It remains unclear however whether such host state-associated variation is a cause or an effect. Our framework additionally facilitated the inference of intra-species population profiles for each individual and our analysis suggested that most individuals harbor multiple strains of each species.

While still far from an exhaustive catalog of strains that may be present across all human gut microbiomes, the framework presented above represents the most comprehensive account of copy number variation in the human gut microbiome to date. It is our hope that this framework and the results presented here will inform future studies of strain-level microbiome composition, demonstrating the extent of functional information that is lost by limiting characterization to the level of species, and prompting further investigation and sequencing of strain-level features. Yet, there are clearly a number of caveats that should be considered in designing future studies. First, our analysis is limited to the detection of variation in gut species for which at least one fully sequenced genome is available. Future studies may benefit from having additional genomes, though notably we

did not detect significantly more variation in clusters for which more reference genomes were available. In addition, our pipeline was designed to detect gene losses or amplifications, but cannot identify gain of genes that are not present in any of the reference genomes included in the genome cluster. Such gain or transfer events may represent an additional substantial source of intra-species variation [151]. Our framework could however further facilitate future efforts to study sequence divergence among duplicated genes, informing our view of neofunctionalization and conservation processes in the microbiome. Notably, in our analysis we focused on detecting high-confidence instances of variation, applying highly-conservative parameters for read alignment and for variability calling. Specifically, we limit our analysis to 'detectable' genome clusters, defined as those with >1x coverage in the sample. Our analysis of a synthetic dataset confirmed that in such clusters copy number estimates can be inferred with 96% accuracy, but that prediction accuracy dropped significantly in genome clusters with lower coverage (Figure C.4B and Appendix C.1). With 13 million reads per sample (the lowest sequencing depth in the cohort analyzed), species that comprise >0.4% of the sample are likely to be considered detectable by our pipeline (while higher sequencing depths of the samples will clearly allow analysis of even rarer species). Future studies may relax some of these parameters or incorporate additional information regarding for example known measures of gene conservation, to detect more subtle variation. Finally, as with most studies relating microbiome composition to function, our analysis relies on the availability of functional gene annotations, which may contain a non-negligible number of erroneous annotations and is clearly incomplete. By considering variation across samples rather than variation from reference genomes, our analysis is largely robust to such annotation inaccuracies and does not report spurious variation when the reference genome is incorrectly annotated. Interestingly, however, many of the variable KCs identified by our analysis had no known functional annotations, and in

fact, variable KCs were much more likely to lack a functional annotation than non-variable KCs. This suggests that much of the variation in gene content among strains detected above has as yet uncharacterized consequences. Combined, these results highlight both the need for additional genome sequences and the importance of continued efforts for characterizing gene function.

Ultimately, however, analyses of intra-species variation in microbial communities such as the one presented above are crucial for understanding the complex relationship between species composition and community-level functional capacity. Looking forward, the insight gained from such analyses may enable the identification of (or engineering of) specific strains or instances of copy number variation that are protective for disease. Notably, current efforts to identify such keystone taxa at the species level have been frustrating on the whole. Additionally, understanding how strain-level variation is distributed throughout human gut populations can not only identify potential novel strains for targeted sequencing, but may help to identify modes of microbial transmission among hosts, and may also help explain why certain host populations experience different metabolic symptoms. Finally, when viewed in a community context, strain-level functional differences will help to explain why some species persist while others decline, with a more nuanced view of the specific functional role and functional dependencies of each organism. Our analysis, quantifiably characterizing strain-level gene variation in the gut microbiome, is an important first step in this direction, and the resulting dataset provides an essential resource for future predictive studies.

5. Concluding Remarks

In this dissertation I have demonstrated the promise of studying microbial consortia under a new framework – one that specifically acknowledges and exploits microbes' unique community-based lifestyle. While straightforward studies of community composition can be limited in their translation to overall community function, the techniques I have presented here focus on bridging this gap, offering three distinct ways to view large-scale metagenomic data through the lens of systems biology. Applying these techniques to a number of datasets has yielded striking insight into the interplay between and among microbes and their host.

First, I demonstrated that simple topological models of community-wide metabolism can be used as a basis for quantitatively characterizing microbiome structure. These models build on comparative studies of community composition by adding important context about functional dependencies between community elements. I showed that these models can reliably discriminate microbial communities associated with different host states, and that the primary discriminatory factors are associated with the microbe-gut interface.

Second, I showed that the configuration of diverse microbiomes from across the mammalian lineage can be decomposed into combinations of fairly discrete gene modules, potentially representing the building blocks of a functional metabolic system. I find that the set of modules present in a given microbiome can be attributed to a combination of both host phylogeny and diet. Previous research had identified individual microbial genes exhibiting differential abundance between mammals in different diet groups, but was unable to translate these differences to full microbial gene profile.

Finally, I conducted a large-scale survey of functional variation between organisms of the same species within the human gut microbiome. These strain-specific differences often go undetected by phylogenetic studies at the species level or higher, yet may represent

important sources of community-dependent adaptation. In addition, this data serves as an important bridge for future studies linking community-wide functional trends back to species of origin. I found that strain variation is widespread, affecting more than 20% of functions in some species, and that specific functional classes are significantly enriched for variation. Interestingly, many of the most variable functions, such as antibiotic resistance and transport, relate to the way a microbe interacts with its environment, demonstrating the importance of context and community in shaping microbial function and adaptation at the organismal level.

These results reinforce the view of a microbiome as a 'super-organism,' a set of potentially diverse organisms whose metabolic potential is defined both by its constituent members and by their interactions. Some of the best support for this view is found in the various underlying biologically-relevant patterns that are readily apparent when applying a systems-level framework, and how these results differ from those reported in the original studies. For example, the original study of 124 human gut microbiomes [130] reported only a difference in the overall diversity of IBD microbiomes compared to healthy individuals, without remarking on any discriminating factors of obesity, and subsequent analyses of the same data have had difficulty in finding any consistent differences among lean and obese phylogenetic profiles [152] . Here however, a systems-level framework reveals consistent differences between the topology of lean and obese microbiome models, which importantly, can be validated in a second independent dataset, and suggest distinct avenues by which these differences may be translated to a functional impact on the host.

Perhaps most notably though, when taken together the systems-level patterns I detect suggest that the environment is tightly intertwined with microbial community structure. In humans, gut-interface genes have the greatest discriminatory abundances between host groups, while differences in gene presence confer host-state-specific topologies with known

environmental relevance. Across gut species, the majority of community-dependent genomic functions are related to environmental sensing and response, while the mammalian gut microbiome may be assembled from co-functional gene groups, whose representation appears to be shaped largely by the environmental influences of host lifestyle and diet.

However, while the models and methods presented here demonstrate the importance of the microbiome-environment relationship, the full extent to which the environment influences composition is still an open question, as a number of other factors certainly play a role and the precise combination of factors is unknown. Similarly, untangling cause from effect remains a major challenge in this work and in most microbiome studies. Does the biochemical gut environment associated with obesity dictate community composition, or is obesity itself a consequence of an altered microbial system? Is community composition a corollary of a particular set of native strains and their unique functional repertoires, or is variation among strains an adaptation to community and environmental context? Other open questions involve microbiome dynamics and robustness. Assessing how quickly changes to community structure take hold, the magnitude of such effects, and how long altered state persists will all be crucial for advancing microbiome manipulation. Creative experimental approaches, detailed host data collection, and more long-term studies will help to tease apart these issues as the field progresses.

Ideally, the ultimate goal of systems-level studies is the integration of these types of questions into a truly predictive model, capable of translating community composition directly to measurable biological outcomes, and capable of predicting the impact of environmental or microbial perturbations on community structure. The modeling frameworks presented here are a first step in this direction, yet moving closer to this ideal will require a number of additional features and a great amount of future work. More complex networks, incorporating both DNA-based gene abundances as well as regulation at the transcript and

protein levels, in addition to reaction rates and dynamics, and detailed information on the host biochemical environment will all be critical additions. Future models may also be better able to bridge the gap between species and gene composition, building on existing species-specific genome-scale metabolic models, and integrating multiple models into a single community-wide stoichiometric model. Finally, I presented here a first exploration of the extent and potential impact of strain variation in a community setting – integrating detailed information about individual community members will be important for accurate prediction of community function as well.

Ultimately, understanding and modeling microbial communities is a complex yet crucially important challenge, with potential implications for some of the most pressing health and environmental issues of our time. Without doubt, there is still much work to be done. Yet, as the field rapidly advances, the hurdle of the next era of analysis will be deriving definitive biological insight from descriptive data-generation studies. The frameworks I present here are a new direction in this vein, setting the stage for a marriage of metagenomic data and systems biology, and providing a first glimpse of the promise to come.

6. References

[1] A. Lewin, A. Wentzel, and S. Valla, “Metagenomics of microbial life in extreme temperature environments.,” *Curr. Opin. Biotechnol.*, vol. 24, no. 3, pp. 516–25, Jun. 2013.

[2] P. D. Schloss and J. Handelsman, “Status of the microbial census.,” *Microbiol. Mol. Biol. Rev.*, vol. 68, no. 4, pp. 686–91, Dec. 2004.

[3] S. L. Lebeis, “The potential for give and take in plant-microbiome relationships.,” *Front. Plant Sci.*, vol. 5, p. 287, Jan. 2014.

[4] M. Hess, A. Sczyrba, R. Egan, T.-W. Kim, H. Chokhawala, G. Schroth, S. Luo, D. S. Clark, F. Chen, T. Zhang, R. I. Mackie, L. A. Pennacchio, S. G. Tringe, A. Visel, T. Woyke, Z. Wang, and E.

M. Rubin, “Metagenomic Discovery of Biomass-Degrading Genes and Genomes from Cow Rumen,” *Science (80-.).*, vol. 331, no. 6016, pp. 463–467, Jan. 2011.

[5] H. A. Gordon and L. Pesti, “The gnotobiotic animal as a tool in the study of host microbial relationships.,” *Bacteriol. Rev.*, vol. 35, no. 4, pp. 390–429, Dec. 1971.

[6] J. L. Round and S. K. Mazmanian, “The gut microbiota shapes intestinal immune responses during health and disease.,” *Nat. Rev. Immunol.*, vol. 9, no. 5, pp. 313–23, May 2009.

[7] M. G. Hale, D. L. Lindsey, and K. M. Hameed, “Gnotobiotic culture of plants and related research,” *Bot. Rev.*, vol. 39, no. 3, pp. 261–273, Jul. 1973.

[8] R. E. Ley, D. A. Peterson, and J. I. Gordon, “Ecological and evolutionary forces shaping microbial diversity in the human intestine.,” *Cell*, vol. 124, no. 4, pp. 837–48, Feb. 2006.

[9] L. V Hooper and J. I. Gordon, “Commensal host-bacterial relationships in the gut.,” *Science*, vol. 292, no. 5519, pp. 1115–8, May 2001.

[10] The Human Microbiome Project Consortium, “Structure, function and diversity of the healthy human microbiome.,” *Nature*, vol. 486, no. 7402, pp. 207–14, Jun. 2012.

[11] H. J. Flint, S. H. Duncan, K. P. Scott, and P. Louis, “Interactions and competition within the microbial community of the human colon: links between diet and health.,” *Environ. Microbiol.*, vol. 9, no. 5, pp. 1101–11, May 2007.

[12] S. K. Hansen, P. B. Rainey, J. A. J. Haagensen, and S. Molin, “Evolution of species interactions in a biofilm community.,” *Nature*, vol. 445, no. 7127, pp. 533–6, Feb. 2007.

[13] E. H. Wintermute and P. A. Silver, “Emergent cooperation in microbial metabolism,” *Mol. Syst. Biol.*, vol. 6, p. 407, Sep. 2010.

[14] J. A. Fuhrman, “Microbial community structure and its functional implications.,” *Nature*, vol. 459, no. 7244, pp. 193–9, May 2009.

[15] A. S. Neish, “Microbes in gastrointestinal health and disease.,” *Gastroenterology*, vol. 136, no. 1, pp. 65–80, Jan. 2009.

[16] J. Qin, R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, D. R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J.-M. Batto, T. Hansen, D. Le Paslier, A. Linneberg, H. B. Nielsen, E. Pelletier, P. Renault, T. Sicheritz-Ponten, K. Turner, H. Zhu, C. Yu, S. Li, M. Jian, Y. Zhou, Y. Li, X. Zhang, S. Li, N. Qin, H. Yang, J. Wang, S. Brunak, J. Doré, F. Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, P. Bork, S. D. Ehrlich, and J. Wang, “A human gut microbial gene catalogue established by metagenomic sequencing.,” *Nature*, vol. 464, no. 7285, pp. 59–65, Mar. 2010.

[17] A. Khoruts, J. Dicksved, J. K. Jansson, and M. J. Sadowsky, “Changes in the composition of the human fecal microbiome after bacteriotherapy for recurrent Clostridium difficile-associated diarrhea.,” *J. Clin. Gastroenterol.*, vol. 44, no. 5, pp. 354–60, 2010.

[18] P. a Vaishampayan, J. V Kuehl, J. L. Froula, J. L. Morgan, H. Ochman, and M. P. Francino, “Comparative metagenomics and population dynamics of the gut microbiota in mother and infant.,” *Genome Biol. Evol.*, vol. 2, pp. 53–66, Jan. 2010.

[19] L. A. David, A. C. Materna, J. Friedman, M. I. Campos-Baptista, M. C. Blackburn, A. Perrotta, S. E. Erdman, and E. J. Alm, “Host lifestyle affects human microbiota on daily timescales,” *Genome Biol.*, vol. 15, no. 7, p. R89, 2014.

[20] P. D. Scanlan, F. Shanahan, and J. R. Marchesi, “Human methanogen diversity and incidence in healthy and diseased colonic groups using mcrA gene analysis.,” *BMC Microbiol.*, vol. 8, no. 1, p. 79, Jan. 2008.

[21] T. A. Clayton, D. Baker, J. C. Lindon, J. R. Everett, and J. K. Nicholson, “Pharmacometabonomic identification of a significant host-microbiome metabolic interaction affecting human drug metabolism.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 34, pp. 14728–33, Aug. 2009.

[22] D. Börnigen, X. C. Morgan, E. A. Franzosa, B. Ren, R. J. Xavier, W. S. Garrett, and C. Huttenhower, “Functional profiling of the gut microbiome in disease-associated inflammation.,” *Genome Med.*, vol. 5, no. 7, p. 65, Jul. 2013.

[23] J. L. Round and S. K. Mazmanian, “The gut microbiota shapes intestinal immune responses during health and disease.,” *Nat. Rev. Immunol.*, vol. 9, no. 5, pp. 313–23, May 2009.

[24] P. D. Schloss and J. Handelsman, “Metagenomics for studying unculturable microorganisms: cutting the Gordian knot.,” *Genome Biol.*, vol. 6, no. 8, p. 229, Jan. 2005.

[25] A. O’Donnell and H. Gorres, “16S rDNA methods in soil microbiology,” *Curr. Opin. Biotechnol.*, vol. 10, no. 3, pp. 225–9, Jun. 1999.

[26] S. Sunagawa, D. R. Mende, G. Zeller, F. Izquierdo-Carrasco, S. A. Berger, J. R. Kultima, L. P. Coelho, M. Arumugam, J. Tap, H. B. Nielsen, S. Rasmussen, S. Brunak, O. Pedersen, F. Guarner, W. M. de Vos, J. Wang, J. Li, J. Doré, S. D. Ehrlich, A. Stamatakis, and P. Bork, “Metagenomic species profiling using universal phylogenetic marker genes.,” *Nat. Methods*, vol. 10, no. 12, pp. 1196–9, Dec. 2013.

[27] J. Peterson, S. Garges, M. Giovanni, P. McInnes, L. Wang, J. A. Schloss, V. Bonazzi, J. E. McEwen, K. A. Wetterstrand, C. Deal, C. C. Baker, V. Di Francesco, T. K. Howcroft, R. W. Karp, R. D. Lunsford, C. R. Wellington, T. Belachew, M. Wright, C. Giblin, H. David, M. Mills, R. Salomon, C. Mullins, B. Akolkar, L. Begg, C. Davis, L. Grandison, M. Humble, J. Khalsa, A. R. Little, H. Peavy, C. Pontzer, M. Portnoy, M. H. Sayre, P. Starke-Reed, S. Zakhari, J. Read, B. Watson, and M. Guyer, “The NIH Human Microbiome Project.,” *Genome Res.*, vol. 19, no. 12, pp. 2317–23, Dec. 2009.

[28] M. Kanehisa, “KEGG: Kyoto Encyclopedia of Genes and Genomes,” *Nucleic Acids Res.*, vol. 28, no. 1, pp. 27–30, Jan. 2000.

[29] R. Overbeek, T. Begley, R. M. Butler, J. V Choudhuri, H.-Y. Chuang, M. Cohoon, V. de Crécy-Lagard, N. Diaz, T. Disz, R. Edwards, M. Fonstein, E. D. Frank, S. Gerdes, E. M. Glass, A. Goesmann, A. Hanson, D. Iwata-Reuyl, R. Jensen, N. Jamshidi, L. Krause, M. Kubal, N. Larsen, B. Linke, A. C. McHardy, F. Meyer, H. Neuweger, G. Olsen, R. Olson, A. Osterman, V. Portnoy, G. D. Pusch, D. A. Rodionov, C. Rückert, J. Steiner, R. Stevens, I. Thiele, O. Vassieva, Y. Ye, O.

Zagnitko, and V. Vonstein, “The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes.,” *Nucleic Acids Res.*, vol. 33, no. 17, pp. 5691–702, Jan. 2005.

[30] R. Caspi, H. Foerster, C. A. Fulcher, P. Kaipa, M. Krummenacker, M. Latendresse, S. Paley, S. Y. Rhee, A. G. Shearer, C. Tissier, T. C. Walk, P. Zhang, and P. D. Karp, “The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases.,” *Nucleic Acids Res.*, vol. 36, no. Database issue, pp. D623–31, Jan. 2008.

[31] K. Kurokawa, T. Itoh, T. Kuwahara, K. Oshima, H. Toh, A. Toyoda, H. Takami, H. Morita, V. K. Sharma, T. P. Srivastava, T. D. Taylor, H. Noguchi, H. Mori, Y. Ogura, D. S. Ehrlich, K. Itoh, T. Takagi, Y. Sakaki, T. Hayashi, and M. Hattori, “Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes.,” *DNA Res.*, vol. 14, no. 4, pp. 169–81, Aug. 2007.

[32] S. G. Tringe, C. von Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur, J. C. Detter, P. Bork, P. Hugenholtz, and E. M. Rubin, “Comparative metagenomics of microbial communities.,” *Science*, vol. 308, no. 5721, pp. 554–7, Apr. 2005.

[33] L. Dethlefsen and D. A. Relman, “Colloquium Paper: Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation,” *Proc. Natl. Acad. Sci.*, vol. 108, no. Supplement_1, pp. 4554–4561, Sep. 2010.

[34] J. J. Faith, N. P. McNulty, F. E. Rey, and J. I. Gordon, “Predicting a human gut microbiota’s response to diet in gnotobiotic mice.,” *Science*, vol. 333, no. 6038, pp. 101–4, Jul. 2011.

[35] B. Rodriguez-Brito, F. Rohwer, and R. A. Edwards, “An application of statistics to comparative metagenomics.,” *BMC Bioinformatics*, vol. 7, no. 1, p. 162, Jan. 2006.

[36] B. Liu and M. Pop, “Statistical methods for comparing the abundances of metabolic pathways in metagenomics,” *Genome Biol.*, vol. 11, no. Suppl 1, p. O7, 2010.

[37] J. Raes, K. U. Foerstner, and P. Bork, “Get the most out of your metagenome: computational analysis of environmental sequence data.,” *Curr. Opin. Microbiol.*, vol. 10, no. 5, pp. 490–8, Oct. 2007.

[38] A. Schwiertz, D. Taras, K. Schäfer, S. Beijer, N. A. Bos, C. Donus, and P. D. Hardt, “Microbiota and SCFA in lean and overweight healthy subjects.,” *Obesity (Silver Spring).*, vol. 18, no. 1, pp. 190–5, Jan. 2010.

[39] F. Armougom, M. Henry, B. Vialettes, D. Raccah, and D. Raoult, “Monitoring bacterial community of human gut microbiota reveals an increase in Lactobacillus in obese patients and Methanogens in anorexic patients.,” *PLoS One*, vol. 4, no. 9, p. e7125, Jan. 2009.

[40] P. J. Turnbaugh, R. E. Ley, M. A. Mahowald, V. Magrini, E. R. Mardis, and J. I. Gordon, “An obesity-associated gut microbiome with increased capacity for energy harvest.,” *Nature*, vol. 444, no. 7122, pp. 1027–31, Dec. 2006.

[41] R. E. Ley, P. J. Turnbaugh, S. Klein, and J. I. Gordon, “Microbial ecology: human gut microbes associated with obesity.,” *Nature*, vol. 444, no. 7122, pp. 1022–3, Dec. 2006.

[42] R. E. Ley, “Obesity and the human microbiome.,” *Curr. Opin. Gastroenterol.*, vol. 26, no. 1, pp. 5–11, Jan. 2010.

[43] P. J. Turnbaugh, M. Hamady, T. Yatsunenko, B. L. Cantarel, A. Duncan, R. E. Ley, M. L. Sogin, W. J. Jones, B. A. Roe, J. P. Affourtit, M. Egholm, B. Henrissat, A. C. Heath, R. Knight, and J. I. Gordon, “A core gut microbiome in obese and lean twins.,” *Nature*, vol. 457, no. 7228, pp. 480–4, Jan. 2009.

[44] H. Kitano, “Systems biology: a brief overview.,” *Science*, vol. 295, no. 5560, pp. 1662–4, Mar. 2002.

[45] I. Thiele, A. Heinken, and R. M. Fleming, “A systems biology approach to studying the role of microbes in human health.,” *Curr. Opin. Biotechnol.*, vol. null, no. null, Oct. 2012.

[46] W. F. Röling, M. Ferrer, and P. N. Golyshin, “Systems approaches to microbial communities and their functioning.,” *Curr. Opin. Biotechnol.*, vol. 21, no. 4, pp. 538–532, Jul. 2010.

[47] R. Levy and E. Borenstein, “Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, no. 31, pp. 12804–9, Jul. 2013.

[48] N. D. Price, J. L. Reed, and B. Ø. Palsson, “Genome-scale models of microbial cells: evaluating the consequences of constraints.,” *Nat. Rev. Microbiol.*, vol. 2, no. 11, pp. 886–97, Nov. 2004.

[49] K. R. Patil and J. Nielsen, “Uncovering transcriptional regulation of metabolism by using metabolic network topology.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 8, pp. 2685–9, Mar. 2005.

[50] S. Freilich, A. Kreimer, E. Borenstein, U. Gophna, R. Sharan, and E. Ruppin, “Decoupling Environment-Dependent and Independent Genetic Robustness across Bacterial Species.,” *PLoS Comput. Biol.*, vol. 6, no. 2, p. e1000690, Jan. 2010.

[51] A. Kreimer, E. Borenstein, U. Gophna, and E. Ruppin, “The evolution of modularity in bacterial metabolic networks.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 19, pp. 6976–81, May 2008.

[52] J. Qin, Y. Li, Z. Cai, S. Li, J. Zhu, F. Zhang, S. Liang, W. Zhang, Y. Guan, D. Shen, Y. Peng, D. Zhang, Z. Jie, W. Wu, Y. Qin, W. Xue, J. Li, L. Han, D. Lu, P. Wu, Y. Dai, X. Sun, Z. Li, A. Tang, S. Zhong, X. Li, W. Chen, R. Xu, M. Wang, Q. Feng, M. Gong, J. Yu, Y. Zhang, M. Zhang, T. Hansen, G. Sanchez, J. Raes, G. Falony, S. Okuda, M. Almeida, E. LeChatelier, P. Renault, N. Pons, J.-M. Batto, Z. Zhang, H. Chen, R. Yang, W. Zheng, S. Li, H. Yang, J. Wang, S. D. Ehrlich, R. Nielsen, O. Pedersen, K. Kristiansen, and J. Wang, “A metagenome-wide association study of gut microbiota in type 2 diabetes.,” *Nature*, vol. 490, no. 7418, pp. 55–60, Oct. 2012.

[53] I. Sharon, S. Bercovici, R. Y. Pinter, and T. Shlomi, “Pathway-based functional analysis of metagenomes.,” *J. Comput. Biol.*, vol. 18, no. 3, pp. 495–505, Mar. 2011.

[54] F. Meyer, D. Paarmann, M. D’Souza, R. Olson, E. M. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening, and R. A. Edwards, “The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes.,” *BMC Bioinformatics*, vol. 9, no. 1, p. 386, Jan. 2008.

[55] V. M. Markowitz, I.-M. A. Chen, K. Palaniappan, K. Chu, E. Szeto, Y. Grechkin, A. Ratner, B. Jacob, J. Huang, P. Williams, M. Huntemann, I. Anderson, K. Mavromatis, N. N. Ivanova, and N. C. Kyrpides, “IMG: the Integrated Microbial Genomes database and comparative analysis system.,” *Nucleic Acids Res.*, vol. 40, no. Database issue, pp. D115–22, Jan. 2012.

[56] S. Abubucker, N. Segata, J. Goll, A. M. Schubert, J. Izard, B. L. Cantarel, B. Rodriguez-Mueller, J. Zucker, M. Thiagarajan, B. Henrissat, O. White, S. T. Kelley, B. Methé, P. D. Schloss, D. Gevers, M. Mitreva, and C. Huttenhower, "Metabolic reconstruction for metagenomic data and its application to the human microbiome.," *PLoS Comput. Biol.*, vol. 8, no. 6, p. e1002358, Jun. 2012.

[57] S. R. Gill, M. Pop, R. T. Deboy, P. B. Eckburg, P. J. Turnbaugh, B. S. Samuel, J. I. Gordon, D. A. Relman, C. M. Fraser-Liggett, and K. E. Nelson, "Metagenomic analysis of the human distal gut microbiome.," *Science*, vol. 312, no. 5778, pp. 1355–9, Jun. 2006.

[58] R. E. Ley, C. A. Lozupone, M. Hamady, R. Knight, and J. I. Gordon, "Worlds within worlds: evolution of the vertebrate gut microbiota.," *Nat. Rev. Microbiol.*, vol. 6, no. 10, pp. 776–88, Oct. 2008.

[59] M. Vijay-Kumar, J. D. Aitken, F. a Carvalho, T. C. Cullender, S. Mwangi, S. Srinivasan, S. V Sitaraman, R. Knight, R. E. Ley, and A. T. Gewirtz, "Metabolic syndrome and altered gut microbiota in mice lacking Toll-like receptor 5.," *Science*, vol. 328, no. 5975, pp. 228–31, Apr. 2010.

[60] P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon, "The human microbiome project.," *Nature*, vol. 449, no. 7164, pp. 804–10, Oct. 2007.

[61] A. L. Hartman, D. M. Lough, D. K. Barupal, O. Fiehn, T. Fishbein, M. Zasloff, and J. A. Eisen, "Human gut microbiome adopts an alternative state following small bowel transplantation.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 40, pp. 17187–92, Oct. 2009.

[62] J. R. White, N. Nagarajan, and M. Pop, "Statistical methods for detecting differentially abundant features in clinical metagenomic samples.," *PLoS Comput. Biol.*, vol. 5, no. 4, p. e1000352, Apr. 2009.

[63] J. Lederberg, "Infectious History," *Science (80-.).*, vol. 288, no. 5464, pp. 287–293, Apr. 2000.

[64] J. I. Gordon and T. R. Klaenhammer, "Colloquium Paper: A rendezvous with our microbes," *Proc. Natl. Acad. Sci.*, vol. 108, no. Supplement_1, pp. 4513–4515, Mar. 2011.

[65] J. Raes and P. Bork, "Molecular eco-systems biology: towards an understanding of community function.," *Nat. Rev. Microbiol.*, vol. 6, no. 9, pp. 693–9, Sep. 2008.

[66] M. A. Oberhardt, B. Ø. Palsson, and J. A. Papin, "Applications of genome-scale metabolic reconstructions.," *Mol. Syst. Biol.*, vol. 5, p. 320, Jan. 2009.

[67] N. Tepper and T. Shlomi, "Predicting metabolic engineering knockout strategies for chemical production: accounting for competing pathways.," *Bioinformatics*, vol. 26, no. 4, pp. 536–43, Feb. 2010.

[68] N. Klitgord and D. Segrè, "Environments that Induce Synthetic Microbial Ecosystems," *PLoS Comput. Biol.*, vol. 6, no. 11, p. e1001002, Nov. 2010.

[69] M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa, "From genomics to chemical genomics: new developments in KEGG.," *Nucleic Acids Res.*, vol. 34, no. Database issue, pp. D354–7, Jan. 2006.

[70] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabási, “The large-scale organization of metabolic networks.,” *Nature*, vol. 407, no. 6804, pp. 651–4, Oct. 2000.

[71] J. Stelling, S. Klamt, K. Bettenbrock, S. Schuster, and E. D. Gilles, “Metabolic network structure determines key aspects of functionality and regulation.,” *Nature*, vol. 420, no. 6912, pp. 190–3, Nov. 2002.

[72] R. Guimerà and L. A. Nunes Amaral, “Functional cartography of complex metabolic networks.,” *Nature*, vol. 433, no. 7028, pp. 895–900, Feb. 2005.

[73] M. C. Palumbo, A. Colosimo, A. Giuliani, and L. Farina, “Functional essentiality from topology features in metabolic networks: a case study in yeast.,” *FEBS Lett.*, vol. 579, no. 21, pp. 4642–6, Aug. 2005.

[74] J. Raymond and D. Segrè, “The effect of oxygen on biochemical networks and the evolution of complex life.,” *Science*, vol. 311, no. 5768, pp. 1764–7, Mar. 2006.

[75] E. Borenstein, M. Kupiec, M. W. Feldman, and E. Ruppin, “Large-scale reconstruction and phylogenetic analysis of metabolic environments.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 38, pp. 14482–7, Sep. 2008.

[76] E. Borenstein and M. W. Feldman, “Topological signatures of species interactions in metabolic networks.,” *J. Comput. Biol.*, vol. 16, no. 2, pp. 191–200, Feb. 2009.

[77] Y. Assenov, F. Ramírez, S.-E. Schelhorn, T. Lengauer, and M. Albrecht, “Computing topological parameters of biological networks.,” *Bioinformatics*, vol. 24, no. 2, pp. 282–4, Jan. 2008.

[78] M. E. J. Newman, “Modularity and community structure in networks.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, no. 23, pp. 8577–82, Jun. 2006.

[79] R. Overbeek, T. Begley, R. M. Butler, J. V Choudhuri, H.-Y. Chuang, M. Cohoon, V. de Crécy-Lagard, N. Diaz, T. Disz, R. Edwards, M. Fonstein, E. D. Frank, S. Gerdes, E. M. Glass, A. Goesmann, A. Hanson, D. Iwata-Reuyl, R. Jensen, N. Jamshidi, L. Krause, M. Kubal, N. Larsen, B. Linke, A. C. McHardy, F. Meyer, H. Neuweger, G. Olsen, R. Olson, A. Osterman, V. Portnoy, G. D. Pusch, D. A. Rodionov, C. Rückert, J. Steiner, R. Stevens, I. Thiele, O. Vassieva, Y. Ye, O. Zagnitko, and V. Vonstein, “The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes.,” *Nucleic Acids Res.*, vol. 33, no. 17, pp. 5691–702, Jan. 2005.

[80] M. L. Green and P. D. Karp, “The outcomes of pathway database computations depend on pathway ontology.,” *Nucleic Acids Res.*, vol. 34, no. 13, pp. 3687–97, Jan. 2006.

[81] S. Freilich, A. Kreimer, E. Borenstein, N. Yosef, R. Sharan, U. Gophna, and E. Ruppin, “Metabolic-network-driven analysis of bacterial ecological strategies.,” *Genome Biol.*, vol. 10, no. 6, p. R61, Jan. 2009.

[82] M. A. Mahowald, F. E. Rey, H. Seedorf, P. J. Turnbaugh, R. S. Fulton, A. Wollam, N. Shah, C. Wang, V. Magrini, R. K. Wilson, B. L. Cantarel, P. M. Coutinho, B. Henrissat, L. W. Crock, A. Russell, N. C. Verberkmoes, R. L. Hettich, and J. I. Gordon, “Characterizing a model human gut microbiota composed of members of its two dominant bacterial phyla.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 14, pp. 5859–64, Apr. 2009.

[83] A. L. Francl, T. Thongaram, and M. J. Miller, “The PTS transporters of Lactobacillus gasseri ATCC 33323.,” *BMC Microbiol.*, vol. 10, no. 1, p. 77, Jan. 2010.

[84] P. J. Turnbaugh, V. K. Ridaura, J. J. Faith, F. E. Rey, R. Knight, and J. I. Gordon, “The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice.,” *Sci. Transl. Med.*, vol. 1, no. 6, p. 6ra14, Nov. 2009.

[85] C.-S. Chen, S. Sullivan, T. Anderson, A. C. Tan, P. J. Alex, S. R. Brant, C. Cuffari, T. M. Bayless, M. V Talor, C. L. Burek, H. Wang, R. Li, L. W. Datta, Y. Wu, R. L. Winslow, H. Zhu, and X. Li, “Identification of novel serological biomarkers for inflammatory bowel disease using Escherichia coli proteome chip.,” *Mol. Cell. Proteomics*, vol. 8, no. 8, pp. 1765–76, Aug. 2009.

[86] G. Kolios, V. Valatas, and S. G. Ward, “Nitric oxide in inflammatory bowel disease: a universal messenger in an unsolved puzzle.,” *Immunology*, vol. 113, no. 4, pp. 427–37, Dec. 2004.

[87] M. Gil-Ortega, P. Stucchi, R. Guzmán-Ruiz, V. Cano, S. Arribas, M. C. González, M. Ruiz-Gayo, M. S. Fernández-Alfonso, and B. Somoza, “Adaptative nitric oxide overproduction in perivascular adipose tissue during early diet-induced obesity.,” *Endocrinology*, vol. 151, no. 7, pp. 3299–306, Jul. 2010.

[88] G.-Y. Yang, S. Taboada, and J. Liao, “Induced nitric oxide synthase as a major player in the oncogenic transformation of inflamed tissue.,” *Methods Mol. Biol.*, vol. 512, pp. 119–56, Jan. 2009.

[89] S. Bernhardsson, P. Gerlee, and L. Lizana, “Structural correlations in bacterial metabolic networks,” *BMC Evol. Biol.*, vol. 11, no. 1, p. 20, 2011.

[90] C. Pál, B. Papp, and M. J. Lercher, “Adaptive evolution of bacterial metabolic networks by horizontal gene transfer.,” *Nat. Genet.*, vol. 37, no. 12, pp. 1372–5, Dec. 2005.

[91] M. Parter, N. Kashtan, and U. Alon, “Environmental variability and modularity of bacterial metabolic networks.,” *BMC Evol. Biol.*, vol. 7, no. 1, p. 169, Jan. 2007.

[92] E. K. Costello, J. I. Gordon, S. M. Secor, and R. Knight, “Postprandial remodeling of the gut microbiota in Burmese pythons.,” *ISME J.*, vol. 4, no. 11, pp. 1375–1385, Jun. 2010.

[93] S. Stolyar, S. Van Dien, K. L. Hillesland, N. Pinel, T. J. Lie, J. A. Leigh, and D. A. Stahl, “Metabolic modeling of a mutualistic microbial community.,” *Mol. Syst. Biol.*, vol. 3, p. 92, Jan. 2007.

[94] S. Freilich, A. Kreimer, I. Meilijson, U. Gophna, R. Sharan, and E. Ruppin, “The large-scale organization of the bacterial network of ecological co-occurrence interactions.,” *Nucleic Acids Res.*, vol. 38, no. 12, pp. 3857–68, Jul. 2010.

[95] T. A. Gianoulis, J. Raes, P. V Patel, R. Bjornson, J. O. Korbel, I. Letunic, T. Yamada, A. Paccanaro, L. J. Jensen, M. Snyder, P. Bork, and M. B. Gerstein, “Quantifying environmental adaptation of metabolic pathways in metagenomics.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 5, pp. 1374–9, Feb. 2009.

[96] M. J. Herrgård, N. Swainston, P. Dobson, W. B. Dunn, K. Y. Arga, M. Arvas, N. Blüthgen, S. Borger, R. Costenoble, M. Heinemann, M. Hucka, N. Le Novère, P. Li, W. Liebermeister, M. L. Mo, A. P. Oliveira, D. Petranovic, S. Pettifer, E. Simeonidis, K. Smallbone, I. Spasić, D. Weichart, R. Brent, D. S. Broomhead, H. V Westerhoff, B. Kirdar, M. Penttilä, E. Klipp, B. Ø. Palsson, U. Sauer, S. G. Oliver, P. Mendes, J. Nielsen, and D. B. Kell, “A consensus yeast metabolic network

reconstruction obtained from a community approach to systems biology.," *Nat. Biotechnol.*, vol. 26, no. 10, pp. 1155–60, Oct. 2008.

[97] I. Thiele and B. Ø. Palsson, "Reconstruction annotation jamborees: a community approach to systems biology.," *Mol. Syst. Biol.*, vol. 6, p. 361, Apr. 2010.

[98] R. E. Ley, M. Hamady, C. Lozupone, P. J. Turnbaugh, R. R. Ramey, J. S. Bircher, M. L. Schlegel, T. A. Tucker, M. D. Schrenzel, R. Knight, and J. I. Gordon, "Evolution of mammals and their gut microbes.," *Science*, vol. 320, no. 5883, pp. 1647–51, Jun. 2008.

[99] F. Delsuc, J. L. Metcalf, L. Wegener Parfrey, S. J. Song, A. González, and R. Knight, "Convergence of gut microbiomes in myrmecophagous mammals.," *Mol. Ecol.*, vol. 23, no. 6, pp. 1301–17, Mar. 2014.

[100] B. D. Muegge, J. Kuczynski, D. Knights, J. C. Clemente, A. González, L. Fontana, B. Henrissat, R. Knight, and J. I. Gordon, "Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans.," *Science*, vol. 332, no. 6032, pp. 970–4, May 2011.

[101] V. Spirin and L. A. Mirny, "Protein complexes and functional modules in molecular networks.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 100, no. 21, pp. 12123–8, Oct. 2003.

[102] O. Cohen, H. Ashkenazy, D. Burstein, and T. Pupko, "Uncovering the co-evolutionary network among prokaryotic genes.," *Bioinformatics*, vol. 28, no. 18, pp. i389–i394, Sep. 2012.

[103] A. Wagner, "Evolutionary constraints permeate large metabolic networks.," *BMC Evol. Biol.*, vol. 9, no. 1, p. 231, Jan. 2009.

[104] R. Carr and E. Borenstein, "Comparative Analysis of Functional Metagenomic Annotation and the Mappability of Short Reads," *PLoS One*, vol. 9, no. 8, p. e105776, Aug. 2014.

[105] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks.," *Genome Res.*, vol. 13, no. 11, pp. 2498–504, Nov. 2003.

[106] E. V Koonin and Y. I. Wolf, "Evolutionary systems biology: links between gene evolution and function.," *Curr. Opin. Biotechnol.*, vol. 17, no. 5, pp. 481–7, Oct. 2006.

[107] R. R. Reisz and J. Fröbisch, "The oldest caseid synapsid from the Late Pennsylvanian of Kansas, and the evolution of herbivory in terrestrial vertebrates.," *PLoS One*, vol. 9, no. 4, p. e94518, Jan. 2014.

[108] O. Deusch, C. O'Flynn, A. Colyer, P. Morris, D. Allaway, P. G. Jones, and K. S. Swanson, "Deep Illumina-based shotgun sequencing reveals dietary effects on the structure and function of the fecal microbiome of growing kittens.," *PLoS One*, vol. 9, no. 7, p. e101021, Jan. 2014.

[109] P. Engel, V. G. Martinson, and N. A. Moran, "Functional diversity within the simple gut microbiota of the honey bee.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 109, no. 27, pp. 11002–7, Jul. 2012.

[110] G. D. Wu, J. Chen, C. Hoffmann, K. Bittinger, Y.-Y. Chen, S. A. Keilbaugh, M. Bewtra, D. Knights, W. A. Walters, R. Knight, R. Sinha, E. Gilroy, K. Gupta, R. Baldassano, L. Nessel, H. Li, F. D. Bushman, and J. D. Lewis, "Linking long-term dietary patterns with gut microbial enterotypes.," *Science*, vol. 334, no. 6052, pp. 105–8, Oct. 2011.

[111] M. Vijay-Kumar, J. D. Aitken, F. A. Carvalho, T. C. Cullender, S. Mwangi, S. Srinivasan, S. V Sitaraman, R. Knight, R. E. Ley, and A. T. Gewirtz, “Metabolic syndrome and altered gut microbiota in mice lacking Toll-like receptor 5.,” *Science*, vol. 328, no. 5975, pp. 228–31, Apr. 2010.

[112] J. M. Kinross, A. W. Darzi, and J. K. Nicholson, “Gut microbiome-host interactions in health and disease.,” *Genome Med.*, vol. 3, no. 3, p. 14, Jan. 2011.

[113] N. Iida, a. Dzutsev, C. a. Stewart, L. Smith, N. Bouladoux, R. a. Weingarten, D. a. Molina, R. Salcedo, T. Back, S. Cramer, R.-M. Dai, H. Kiu, M. Cardone, S. Naik, a. K. Patri, E. Wang, F. M. Marincola, K. M. Frank, Y. Belkaid, G. Trinchieri, and R. S. Goldszmid, “Commensal Bacteria Control Cancer Response to Therapy by Modulating the Tumor Microenvironment,” *Science (80-.).*, vol. 342, no. 6161, pp. 967–970, Nov. 2013.

[114] D. N. Frank, A. L. St Amand, R. A. Feldman, E. C. Boedeker, N. Harpaz, and N. R. Pace, “Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 104, no. 34, pp. 13780–5, Aug. 2007.

[115] N. Larsen, F. K. Vogensen, F. W. J. van den Berg, D. S. Nielsen, A. S. Andreasen, B. K. Pedersen, W. A. Al-Soud, S. J. Sørensen, L. H. Hansen, and M. Jakobsen, “Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults.,” *PLoS One*, vol. 5, no. 2, p. e9085, Jan. 2010.

[116] L. R. Hoffman, C. E. Pope, H. S. Hayden, S. Heltshe, R. Levy, S. McNamara, M. A. Jacobs, L. Rohmer, M. Radey, B. W. Ramsey, M. J. Brittnacher, E. Borenstein, and S. I. Miller, “Escherichia coli dysbiosis correlates with gastrointestinal dysfunction in children with cystic fibrosis.,” *Clin. Infect. Dis.*, vol. 58, no. 3, pp. 396–9, Feb. 2014.

[117] N. Salama, K. Guillemin, T. K. McDaniel, G. Sherlock, L. Tompkins, and S. Falkow, “A whole-genome microarray reveals genetic diversity among Helicobacter pylori strains.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 97, no. 26, pp. 14668–73, Dec. 2000.

[118] S. R. Gill, D. E. Fouts, G. L. Archer, E. F. Mongodin, R. T. Deboy, J. Ravel, I. T. Paulsen, J. F. Kolonay, L. Brinkac, M. Beanan, R. J. Dodson, S. C. Daugherty, R. Madupu, S. V Angiuoli, A. S. Durkin, D. H. Haft, J. Vamathevan, H. Khouri, T. Utterback, C. Lee, G. Dimitrov, L. Jiang, H. Qin, J. Weidman, K. Tran, K. Kang, I. R. Hance, K. E. Nelson, and C. M. Fraser, “Insights on evolution of virulence and resistance from the complete genome analysis of an early methicillin-resistant Staphylococcus aureus strain and a biofilm-producing methicillin-resistant Staphylococcus epidermidis strain.,” *J. Bacteriol.*, vol. 187, no. 7, pp. 2426–38, Apr. 2005.

[119] M. Solheim, A. Aakra, L. G. Snipen, D. A. Brede, and I. F. Nes, “Comparative genomics of Enterococcus faecalis from healthy Norwegian infants.,” *BMC Genomics*, vol. 10, p. 194, Jan. 2009.

[120] P. Zunino, C. Piccini, and C. Legnani-Fajardo, “Flagellate and non-flagellate Proteus mirabilis in the development of experimental urinary tract infection,” *Microbial Pathogenesis*, vol. 16, no. 5. pp. 379–385, 1994.

[121] R. J. Siezen, V. A. Tzeneva, A. Castioni, M. Wels, H. T. K. Phan, J. L. W. Rademaker, M. J. C. Starrenburg, M. Kleerebezem, D. Molenaar, and J. E. T. van Hylckama Vlieg, “Phenotypic and

genomic diversity of Lactobacillus plantarum strains isolated from various environmental niches.," *Environ. Microbiol.*, vol. 12, no. 3, pp. 758–73, Mar. 2010.

[122] L. Kraal, S. Abubucker, K. Kota, M. A. Fischbach, and M. Mitreva, "The prevalence of species and strains in the human microbiome: a resource for experimental efforts.," *PLoS One*, vol. 9, no. 5, p. e97279, Jan. 2014.

[123] B. D. Muegge, J. Kuczynski, D. Knights, J. C. Clemente, A. González, L. Fontana, B. Henrissat, R. Knight, and J. I. Gordon, "Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans.," *Science*, vol. 332, no. 6032, pp. 970–4, May 2011.

[124] S. Schloissnig, M. Arumugam, S. Sunagawa, M. Mitreva, J. Tap, A. Zhu, A. Waller, D. R. Mende, J. R. Kultima, J. Martin, K. Kota, S. R. Sunyaev, G. M. Weinstock, and P. Bork, "Genomic variation landscape of the human gut microbiome.," *Nature*, vol. 493, no. 7430, pp. 45–50, Jan. 2013.

[125] M. S. Fitzsimons, M. Novotny, C.-C. Lo, A. E. K. Dichosa, J. L. Yee-Greenbaum, J. P. Snook, W. Gu, O. Chertkov, K. W. Davenport, K. McMurry, K. G. Reitenga, A. R. Daughton, J. He, S. L. Johnson, C. D. Gleasner, P. L. Wills, B. Parson-Quintana, P. S. Chain, J. C. Detter, R. S. Lasken, and C. S. Han, "Nearly finished genomes produced using gel microdroplet culturing reveal substantial intraspecies genomic diversity within the human microbiome.," *Genome Res.*, vol. 23, no. 5, pp. 878–88, May 2013.

[126] E. E. Hansen, C. A. Lozupone, F. E. Rey, M. Wu, J. L. Guruge, A. Narra, J. Goodfellow, J. R. Zaneveld, D. T. McDonald, J. A. Goodrich, A. C. Heath, R. Knight, and J. I. Gordon, "Pan-genome of the dominant human gut-associated archaeon, Methanobrevibacter smithii, studied in twins.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108 Suppl , pp. 4599–606, Mar. 2011.

[127] C. T. Brown, I. Sharon, B. C. Thomas, C. J. Castelle, M. J. Morowitz, and J. F. Banfield, "Genome resolved analysis of a premature infant gut microbial community reveals a Varibaculum cambriense genome and a shift towards fermentation-based metabolism during the third week of life.," *Microbiome*, vol. 1, no. 1, p. 30, Jan. 2013.

[128] I. Sharon, M. J. Morowitz, B. C. Thomas, E. K. Costello, D. A. Relman, and J. F. Banfield, "Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization.," *Genome Res.*, vol. 23, no. 1, pp. 111–20, Jan. 2013.

[129] M. J. Morowitz, V. J. Denef, E. K. Costello, B. C. Thomas, V. Poroyko, D. A. Relman, and J. F. Banfield, "Strain-resolved community genomic analysis of gut microbial colonization in a premature infant.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108, no. 3, pp. 1128–1133, Dec. 2010.

[130] J. Qin, R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, D. R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J.-M. Batto, T. Hansen, D. Le Paslier, A. Linneberg, H. B. Nielsen, E. Pelletier, P. Renault, T. Sicheritz-Ponten, K. Turner, H. Zhu, C. Yu, S. Li, M. Jian, Y. Zhou, Y. Li, X. Zhang, S. Li, N. Qin, H. Yang, J. Wang, S. Brunak, J. Doré, F. Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, P. Bork, S. D. Ehrlich, and J. Wang, "A human gut microbial gene catalogue established by metagenomic sequencing.," *Nature*, vol. 464, no. 7285, pp. 59–65, Mar. 2010.

[131] Y. Kodama, J. Mashima, E. Kaminuma, T. Gojobori, O. Ogasawara, T. Takagi, K. Okubo, and Y. Nakamura, "The DNA Data Bank of Japan launches a new resource, the DDBJ Omics Archive of

functional genomics experiments.," *Nucleic Acids Res.*, vol. 40, no. Database issue, pp. D38–42, Jan. 2012.

[132] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin, "The Sequence Alignment/Map format and SAMtools.," *Bioinformatics*, vol. 25, no. 16, pp. 2078–9, Aug. 2009.

[133] R. V. Sonti and J. R. Roth, "Role of Gene Duplications in the Adaptation of Salmonella typhimurium to Growth on Limiting Carbon Sources," *Genetics*, vol. 123, no. 1, pp. 19–28, Sep. 1989.

[134] D. Gevers, K. Vandepoele, C. Simillion, and Y. Van de Peer, "Gene duplication and biased functional retention of paralogs in bacterial genomes," *Trends Microbiol.*, vol. 12, no. 4, pp. 148–154, Apr. 2004.

[135] E. Heikkinen, T. Kallonen, L. Saarinen, R. Sara, A. J. King, F. R. Mooi, J. T. Soini, J. Mertsola, and Q. He, "Comparative genomics of Bordetella pertussis reveals progressive gene loss in Finnish strains.," *PLoS One*, vol. 2, no. 9, p. e904, Jan. 2007.

[136] M.-C. Lee and C. J. Marx, "Repeated, selection-driven genome reduction of accessory genes in experimental populations.," *PLoS Genet.*, vol. 8, no. 5, p. e1002651, Jan. 2012.

[137] X. Han, R. M. Kennan, J. K. Davies, L. A. Reddacliff, O. P. Dhungyel, R. J. Whittington, L. Turnbull, C. B. Whitchurch, and J. I. Rood, "Twitching motility is essential for virulence in Dichelobacter nodosus.," *J. Bacteriol.*, vol. 190, no. 9, pp. 3323–35, May 2008.

[138] A. Al Mamun, A. Tominaga, and M. Enomoto, "Cloning and characterization of the region III flagellar operons of the four Shigella subgroups: genetic defects that cause loss of flagella of Shigella boydii and Shigella sonnei," *J. Bacteriol.*, vol. 179, no. 14, pp. 4493–4500, Jul. 1997.

[139] K. Borziak, A. D. Fleetwood, and I. B. Zhulin, "Chemoreceptor gene loss and acquisition via horizontal gene transfer in Escherichia coli.," *J. Bacteriol.*, vol. 195, no. 16, pp. 3596–602, Aug. 2013.

[140] B. A. Neville, P. O. Sheridan, H. M. B. Harris, S. Coughlan, H. J. Flint, S. H. Duncan, I. B. Jeffery, M. J. Claesson, R. P. Ross, K. P. Scott, and P. W. O'Toole, "Pro-inflammatory flagellin proteins of prevalent motile commensal bacteria are variably abundant in the intestinal microbiome of elderly humans.," *PLoS One*, vol. 8, no. 7, p. e68919, Jan. 2013.

[141] H. Daniel, A. Moghaddas Gholami, D. Berry, C. Desmarchelier, H. Hahne, G. Loh, S. Mondot, P. Lepage, M. Rothballer, A. Walker, C. Böhm, M. Wenning, M. Wagner, M. Blaut, P. Schmitt-Kopplin, B. Kuster, D. Haller, and T. Clavel, "High-fat diet alters gut microbiota physiology in mice.," *ISME J.*, vol. 8, no. 2, pp. 295–308, Feb. 2014.

[142] S. Furukawa, T. Fujita, M. Shimabukuro, M. Iwaki, Y. Yamada, Y. Nakajima, O. Nakayama, M. Makishima, M. Matsuda, and I. Shimomura, "Increased oxidative stress in obesity and its impact on metabolic syndrome.," *J. Clin. Invest.*, vol. 114, no. 12, pp. 1752–61, Dec. 2004.

[143] F. M. Sohet, A. M. Neyrinck, B. D. Pachikian, F. C. de Backer, L. B. Bindels, P. Niklowitz, T. Menke, P. D. Cani, and N. M. Delzenne, "Coenzyme Q10 supplementation lowers hepatic oxidative stress and inflammation associated with diet-induced obesity in mice.," *Biochem. Pharmacol.*, vol. 78, no. 11, pp. 1391–400, Dec. 2009.

[144] P. H. Langsjoen and A. M. Langsjoen, “Comparison study of plasma coenzyme Q 10 levels in healthy subjects supplemented with ubiquinol versus ubiquinone,” *Clin. Pharmacol. Drug Dev.*, vol. 3, no. 1, pp. 13–17, Jan. 2014.

[145] K. Machiels, M. Joossens, J. Sabino, V. De Preter, I. Arijs, V. Eeckhaut, V. Ballet, K. Claes, F. Van Immerseel, K. Verbeke, M. Ferrante, J. Verhaegen, P. Rutgeerts, and S. Vermeire, “A decrease of the butyrate-producing species Roseburia hominis and Faecalibacterium prausnitzii defines dysbiosis in patients with ulcerative colitis.,” *Gut*, vol. 63, no. 8, pp. 1275–83, Aug. 2014.

[146] H. Sokol, B. Pigneur, L. Watterlot, O. Lakhdari, L. G. Bermúdez-Humarán, J.-J. Gratadoux, S. Blugeon, C. Bridonneau, J.-P. Furet, G. Corthier, C. Grangette, N. Vasquez, P. Pochart, G. Trugnan, G. Thomas, H. M. Blottière, J. Doré, P. Marteau, P. Seksik, and P. Langella, “Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 43, pp. 16731–6, Oct. 2008.

[147] S. Saitoh, S. Noda, Y. Aiba, A. Takagi, M. Sakamoto, Y. Benno, and Y. Koga, “Bacteroides ovatus as the Predominant Commensal Intestinal Microbe Causing a Systemic Antibody Response in Inflammatory Bowel Disease,” *Clin. Vaccine Immunol.*, vol. 9, no. 1, pp. 54–59, Jan. 2002.

[148] M. G. I. Langille, J. Zaneveld, J. G. Caporaso, D. McDonald, D. Knights, J. A. Reyes, J. C. Clemente, D. E. Burkepile, R. L. Vega Thurber, R. Knight, R. G. Beiko, and C. Huttenhower, “Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences.,” *Nat. Biotechnol.*, vol. 31, no. 9, pp. 814–21, Sep. 2013.

[149] J. R. Zaneveld, C. Lozupone, J. I. Gordon, and R. Knight, “Ribosomal RNA diversity predicts genome diversity in gut bacteria and their relatives.,” *Nucleic Acids Res.*, vol. 38, no. 12, pp. 3869–79, Jul. 2010.

[150] B. J. Shapiro, J. Friedman, O. X. Cordero, S. P. Preheim, S. C. Timberlake, G. Szabó, M. F. Polz, and E. J. Alm, “Population genomics of early events in the ecological differentiation of bacteria.,” *Science*, vol. 336, no. 6077, pp. 48–51, Apr. 2012.

[151] C. S. Smillie, M. B. Smith, J. Friedman, O. X. Cordero, L. A. David, and E. J. Alm, “Ecology drives a global network of gene exchange connecting the human microbiome.,” *Nature*, vol. 480, no. 7376, pp. 241–4, Dec. 2011.

[152] M. M. Finucane, T. J. Sharpton, T. J. Laurent, and K. S. Pollard, “A taxonomic signature of obesity in the microbiome? Getting to the guts of the matter.,” *PLoS One*, vol. 9, no. 1, p. e84689, Jan. 2014.

[153] Z. Ning, A. J. Cox, and J. C. Mullikin, “SSAHA: a fast search method for large DNA databases.,” *Genome Res.*, vol. 11, no. 10, pp. 1725–9, Oct. 2001.

[154] W. Xie, F. Wang, L. Guo, Z. Chen, S. M. Sievert, J. Meng, G. Huang, Y. Li, Q. Yan, S. Wu, X. Wang, S. Chen, G. He, X. Xiao, and A. Xu, “Comparative metagenomics of microbial communities inhabiting deep-sea hydrothermal vent chimneys with contrasting chemistries.,” *ISME J.*, vol. 5, no. 3, pp. 414–26, Mar. 2011.

[155] D. G. Ahren and C. A. Ouzounis, “Robustness of metabolic map reconstruction.,” *J. Bioinform. Comput. Biol.*, vol. 2, no. 3, pp. 589–93, Sep. 2004.

[156] H. Ma and A.-P. Zeng, “Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms.,” *Bioinformatics*, vol. 19, no. 2, pp. 270–7, Jan. 2003.

[157] A. Wagner and D. A. Fell, “The small world inside large metabolic networks.,” *Proc. Biol. Sci.*, vol. 268, no. 1478, pp. 1803–10, Sep. 2001.

[158] M. Huss and P. Holme, “Currency and commodity metabolites: their identification and relation to the modularity of metabolic networks.,” *IET Syst. Biol.*, vol. 1, no. 5, pp. 280–5, Sep. 2007.

[159] C. S. Henry, M. DeJongh, A. A. Best, P. M. Frybarger, B. Linsay, and R. L. Stevens, “High-throughput generation, optimization and analysis of genome-scale metabolic models.,” *Nat. Biotechnol.*, vol. 28, no. 9, pp. 977–82, Sep. 2010.

[160] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, “Network motifs: simple building blocks of complex networks.,” *Science*, vol. 298, no. 5594, pp. 824–7, Oct. 2002.

[161] V. M. Markowitz, N. N. Ivanova, E. Szeto, K. Palaniappan, K. Chu, D. Dalevi, I.-M. A. Chen, Y. Grechkin, I. Dubchak, I. Anderson, A. Lykidis, K. Mavromatis, P. Hugenholtz, and N. C. Kyrpides, “IMG/M: a data management and analysis system for metagenomes.,” *Nucleic Acids Res.*, vol. 36, no. Database issue, pp. D534–8, Jan. 2008.

[162] B. D. Wallace, H. Wang, K. T. Lane, J. E. Scott, J. Orans, J. S. Koo, M. Venkatesh, C. Jobin, L.-A. Yeh, S. Mani, and M. R. Redinbo, “Alleviating cancer drug toxicity by inhibiting a bacterial enzyme.,” *Science*, vol. 330, no. 6005, pp. 831–5, Nov. 2010.

[163] Z. Wang, E. Klipfell, B. J. Bennett, R. Koeth, B. S. Levison, B. Dugar, A. E. Feldstein, E. B. Britt, X. Fu, Y.-M. Chung, Y. Wu, P. Schauer, J. D. Smith, H. Allayee, W. H. W. Tang, J. A. DiDonato, A. J. Lusis, and S. L. Hazen, “Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease.,” *Nature*, vol. 472, no. 7341, pp. 57–63, Apr. 2011.

[164] T. A. Clayton, D. Baker, J. C. Lindon, J. R. Everett, and J. K. Nicholson, “Pharmacometabonomic identification of a significant host-microbiome metabolic interaction affecting human drug metabolism.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 34, pp. 14728–33, Aug. 2009.

[165] M. Kircher, U. Stenzel, and J. Kelso, “Improved base calling for the Illumina Genome Analyzer using machine learning strategies.,” *Genome Biol.*, vol. 10, no. 8, p. R83, Jan. 2009.

Appendix A: Supplementary material for Chapter 2

A.1 Supplementary Methods

A.1.1 Datasets

For the fecal microbiomes from unrelated individuals, metagenomic shotgun sequencing reads, predicted genes from the assembled scaffolds, and KEGG annotations [69] were downloaded from *http://www.bork.embl.de/~arumugam/Qin_et_al_2010/* [16]. In order to obtain abundance information, the sequencing reads from each sample were mapped onto the non-redundant gene catalog with SSAHA2 (parameters: -output cigar -score 20 -best 1 -solexa -kmer 12 -skip 12) [153]. Metagenomic shotgun sequencing reads from a smaller set of twin-mother trios [43] were re-analyzed separately, including samples from 9 lean/overweight and 9 obese individuals. All sequences were mapped onto the KEGG database version 52 (BLASTX e-value<10^{-5}, %identity>50, score>50) (see [43] for a detailed discussion of the parameters used). Sequences were annotated with all KOs (KEGG orthologous groups) of the top KO-associated match among the top 100 matches. Sequences with multiple top KO-associated matches with the same e-value were annotated with the union set of KOs. Following the rationale of [16], we imposed a threshold of 2 reads to allow the inclusion of rare genes; all KO abundances below this threshold were set to zero. For both datasets, counts were normalized within each sample to represent the relative abundance of each enzyme (KO) in each sample, thereby accounting for differences in sampling depth. Mean normalized enzyme counts across the two datasets were significantly correlated (R=0.86; Spearman correlation test).

A.1.2 Analysis of Data from Turnbaugh et al

In addition to the analysis of the 124 deeply sequenced microbiomes [16], we performed a similar analysis of 18 microbiomes from mother-twin trios characterized as lean/overweight (n=9) or obese (n=9) [43]. Of the 1195 enzymes in the community-level network constructed from these microbiomes, 120 were enriched in obese microbiomes (odds ratio>2) and 177 were depleted (odds ratio<0.5). 1159 (97%) of the enzymes were also present in the network constructed from the 124 samples of [16]. Tests of association (Table A.3) between differential abundance score and network topology revealed that

differential abundance was negatively correlated with centrality ($p<9.7x10^{-8}$; Spearman correlation test). The centrality scores of obesity-associated enzymes in this dataset are significantly lower than the centrality scores of enzymes not associated with obesity ($p<0.0002$; Wilcoxon rank-sum test; $p<0.04$ and $p<0.0006$ respectively when considering obesity-enriched and obesity-depleted enzymes separately). These enzymes are significantly over-represented in the peripheral tier of the network ($p<0.0012$; Hypergeometric test): 30.4%, 24.8%, and 19.3% of the enzymes in the peripheral, intermediate and central tiers respectively are associated with obesity.

A.1.3 Communities as Supra-organisms

Microbial communities have often been described as supra-organisms, ignoring the individual species comprising each community, and treating the community as a single adaptive organism that functions in a given environment. While there are several reasons to assume that for a large scale analysis the partition of a metabolic pathway among the various species and the compartmentalization of metabolic processes can be ignored, this assumption is often the outcome of necessity; reliable methods for decomposing complex metagenomic samples into species-specific data are currently lacking. Comparative metagenomic analysis, by definition, takes a similar view, assuming that the set of genes found in a metagenome reflects community-level metabolic capacity. In this study, we extend this approach, putting microbial communities on an equal footing with single species and treating the entire community as a single, independent biological system [64]. This approach allows us to ask questions about the organization of the community, its function, and its interaction with the host, and to address these questions using in silico modeling techniques and systems-level methods originally developed for studying single organisms. Specifically, we can project variation in gene content onto a community-level metabolic network and identify systems-level variations that cannot be revealed with traditional comparative analysis.

A.1.4 Choice of Enrichment Metric

The selection of a suitable measure of gene or pathway enrichment is a non-trivial component of comparative metagenomic analyses, since statistical results can be heavily influenced by missing data,

sample size, data magnitude, normalization technique, or assumptions about the distribution of values. Following the example of recent metagenomic studies [43], [57], [154] we chose the odds ratio as the most appropriate measure of enzyme enrichment. The odds ratio measure has a number of important benefits. It has a straightforward *interpretation*, conveying the likelihood of observing a given enzyme in the sample set relative to a comparison data set. It provides a *continuous*, *unit-free* measure of comparative abundance, without assuming any specific *distribution* or *scale* within the data values. This is especially important when assessing hundreds to thousands of enzymes, each with a considerably different range of expected abundances. See also "*Alternative Enrichment Metrics*" below.

A.1.5 Sensitivity of Results to Missing or Erroneous Annotation Data:

We examined the effect of missing or erroneous metabolic annotation data on the calculated centrality scores and on the observed correlation between centrality and differential abundance in obesity using a simulation analysis. Specifically, we considered the available metabolic annotations from the entire KEGG database as a reference ''complete'' dataset and examined the effect of either deleting a varying fraction of all enzymatic annotations in this dataset or rewiring some of the derived links between enzymes. For each deletion level we generated 100 perturbed datasets (i.e., a dataset in which a given percentage of all enzymatic annotations are deleted at random), used these perturbed datasets to construct metabolic networks, and examined the centrality scores and centrality-related results in these networks. As metabolic network construction that is based on as low as 50% gene coverage may still detect 70% of network reactions [155], we analyzed the effect of deleting up to 30% of the enzymatic annotations included in KEGG.

We first examined the effect of such deletions on the calculated centrality values used in our analysis. As is evident from Fig. A.4, the centrality values of the remaining enzymes are hardly affected, even when 30% of the annotations in the KEGG database are deleted. More importantly, as our analysis concerns the relative centrality of the various enzymes (rather than their absolute values), we find that the centrality values obtained for the perturbed datasets are markedly correlated with the centrality values in the original full dataset (R=0.998, 0.996, 0.984, and 0.97 after deleting 5%, 10%, 20%, and 30% of the annotations respectively; Spearman correlation test, $p<10^{-324}$ in all cases). Rewiring 10% of the links

between enzymes (to simulate erroneous annotations) similarly yields highly correlated centrality values ($R=0.79$, $p<10^{-324}$).

We additionally checked whether the correlation between centrality and obesity-associated differential abundance that was observed in the complete KEGG-based network still holds in the perturbed, incomplete networks. We found that in *all* 100 networks based on either 5%, 10% and 20% deleted annotations and in 98 of the 100 networks based on 30% deleted annotations, this correlation analysis still yields a significant result. Figs. A.5*A-B* further demonstrate the relatively little effect incomplete annotation may have on the markedly different centrality scores of obesity-associated vs. non-associated enzymes and on the over-representation of obesity-associated enzymes in the periphery of the network. Similarly, all 100 perturbed networks in which 10% of the links were rewired exhibit a significant correlation between centrality and obesity-associated differential abundance.

A.1.6 Alternative Enrichment Metrics

While we believe that the odds ratio test is the most appropriate metric for our data (as described above, in the section *"Choice of Enrichment Metric"*), we confirmed that our main results are not an artifact of the odds ratio metric specifically and that these results hold under several alternative strategies for identifying enzymes associated with a given host state. These strategies intended to capture various aspects of the possible association between enzyme abundances and the state of the host.

First, we examined the effect of augmenting the odds radio criterion with statistical tests aimed at identifying enzymes with a consistently different abundance profile between the two host states (e.g., obese vs. lean). One such test is based on shuffling the sample labels and is described in more detail in the section *"Robustness to Noise in Read Count Data"* below. In a second test, we alternatively used a Wilcoxon rank-sum test to further limit enzymes that appear to be enriched or depleted according to the odds-ratio test to those that exhibit a significantly ($p<0.05$) different set of abundances values in obese or IBD according to the ranksum test.

Second, we used a fundamentally different test, based solely on the presence or absence of enzymes in the various samples rather than on their abundances. Specifically, we identified enzymes that are present in a significantly high or a significantly low number of samples from a given host state

considering the total number of samples in which they are present. A Hypergeometric test was used to calculated over- or under-representation with a threshold of $p<0.05$.

Next, we used a test based on the *difference* in enzyme abundance, rather than on the ratio. In order to avoid biasing this measure toward the most abundant enzymes overall, we first mapped enzyme abundances to a uniform distribution by ranking the enzymes within each sample from most abundant to least abundant. We then measured the difference between the mean rank of each enzyme in obese samples and its mean rank in lean-healthy samples. Since no clear or intuitive threshold (analogous to the two-fold threshold commonly used in the odds-ratio test) exists for rank difference, here we defined the enzymes with the 10% highest calculated differences and the 10% lowest differences as enriched and depleted enzymes respectively.

Finally, we examined the divergence in the overall distribution of enzyme abundance between the two host states. To this end, we use the Jensen–Shannon divergence measure to quantify the similarity between the distributions of abundance values of a given enzymes in obese vs. lean samples. We again used rank-normalized enzyme abundances to avoid bias. We then bin the rank values associated with each enzyme across the various samples to obtain a distribution profile. Since Jensen–Shannon divergence provides an absolute (rather than signed) measure and does not separate enriched from deplete enzymes, here we defined all enzymes with the 20% most divergent distributions as obesity-associated (using only the 10% most divergent enzymes did not qualitatively change the results).

We find that applying any of these metrics to identify enzymes associated with obesity does not qualitatively change the patterns reported in the main text. Specifically, the strong link between obesity-associated enzymes and centrality is still observed under all these metrics. The pertaining p-values and additional details can be found in Table S3.

A.1.7 Analysis of Non Transport-Related Enzymes

As transport enzymes are likely to be found at the periphery of the network, we further validated that the correlation between differential abundance and centrality is not a product of the over-representation of transport enzymes in obese microbiomes. Omitting all enzymes annotated with a transport-related function (i.e., KEGG BRITE classes ‘Membrane Transport’ and ‘Transport and

Catabolism') and repeating the analysis outlined above, we found that the differential abundance score of the remaining enzymes is still negatively correlated with their centrality ($R=-0.16$, $p<6.6\times10^{-10}$ [obesity]; $R=-0.14$, $p<2.8\times10^{-8}$ [IBD]; Spearman correlation test). Similarly, the centrality of non-transport-related obesity-associated enzymes is significantly lower than centrality scores of enzymes not associated with obesity ($p<1.6\times10^{-5}$ [obesity]; $p<3.2\times10^{-5}$ [IBD]; Wilcoxon rank-sum test) and such non-transport obesity-associated enzymes are significantly over-represented in the peripheral tier of the network ($p<3.1\times10^{-5}$ [obesity]; $p<1.7\times10^{-4}$ [IBD]; Hypergeometric enrichment test).

A.1.8 Robustness to Noise in Read Count Data

Clearly, shotgun metagenomic data is extremely noisy and often represents only a sparse sample of the genomic material found in the microbiome. This noise induces a non-negligible level of inaccuracy in read count and gene abundance data, and consequently in our differential abundance estimates. Such inaccuracies are especially severe in genes with an overall low abundance where, for example, a two-fold difference (e.g., from 1 to 2 reads) most probably represents a sampling error rather than a real signal of association with a given host state.

To confirm that the findings presented in the main text are not an artifact stemming from such low count or noisy data, we used two assays. First we confirmed that the link between obesity-associated or IBD-associated enzymes and centrality holds when the analysis is limited to enzymes with a substantial read count. The analysis presented in the main text follows the rationale presented in [16] and imposes a threshold of 2 reads (see also the "Datasets" section above). Here, we further validated that using a threshold of 5, 25, or 50 reads does not qualitatively affect the results reported in the main text. Specifically, using any of these threshold values, we found that differential abundance scores are still negatively correlated with centrality, centrality scores of obesity- or IBD-associated enzymes are significantly lower than centrality scores of non-associated enzymes, and these enzymes are significantly over-represented in the peripheral tier of the network (see Table S3 for details and p-values).

Second, we applied a measure of significance to our odds ratio score by shuffling the sample labels 1,000 times and recalculating the odds ratio of each enzyme using these shuffled datasets. A p-value was assigned to each enzyme representing the fraction of times an odds ratio more extreme than

the real value was obtained. Enzymes with low p-values therefore represent enzymes that are *consistently* enriched or depleted in obese or IBD samples and are more likely to capture genuine association with the host state. In contrast, enzymes that appear to be differentially abundant only due to a few exceptionally high (or low) counts in some samples (e.g., due to noise) will fail such a test. We classified enzymes with an odds ratio >2 and a p-value < 0.05 as *consistently enriched,* and enzymes with an odds ratio <0.5 and a p-value < 0.05 as *consistently depleted*. Repeating our analysis using this more stringent criterion for association with obesity or IBD still demonstrated a significant correlation between such consistently associated enzymes and centrality (see Table S3 and main text).

Finally, it should also be noted that the correlation between centrality and differential abundance scores was observed in two independent datasets: An Illumina-derived data from 124 individuals and a 454 FLX–derived data from 18 individuals (see Table S3). These two datasets represents fundamentally different metagenomic technologies and different noise profiles.

A.1.9 Robustness to Alternative Network Construction Methods

The construction of a network that accurately depicts the relationships between metabolic enzymes operating in the gut microbiome is central to our analysis. The network generated based on the metagenome content represents our best available approximation of the gut microbiome metabolism and we therefore base our analysis in the main text on this network. However, this network is clearly limited by sampling depth and detectable enzymes and may accordingly be incomplete and inaccurate. Moreover, the network construction strategy and the specific annotation framework used may further affect the topology of the resulting network. We therefore validate here that our findings are not an artifact of the specific network construction method or annotation system. To this end, we generated networks using a number of alternative strategies and examine the link between host-associated enzymes and topology in these alternative networks.

We first constructed networks based on extensions or modifications of the KEGG-based enzyme annotation. The first of these networks was constructed from all enzymes in the KEGG database rather than the subset of enzymes found in the metagenome data. This network accounts for enzymes that do not appear in our data-derived network due to sampling bias or low abundance. A second alternative

network was constructed by omitting currency metabolites from the network wiring scheme. Currency metabolites (such as ATP or H_2O) are those metabolites which are common to a very large number of reactions and could therefore induce overly-dense clusters of connected enzymes with limited biological meaning. The omission of such metabolites is a common practice in the construction of many metabolite-based networks [156]. We compiled a list of currency metabolites (including ATP, ADP, CO_2, H+, H_2, H_2CO_3, H_2O, H_2O_2, H_2S, NAD+, NADH, NH_3, Nitrate, Nitric oxide, Nitrite, O_2, Phosphate, Pyrophosphate, Sulfate, and Sulfite) from a search of relevant literature [156]–[158] and constructed a network in which these metabolites were ignored when generating edges.

We next examined the effect of reaction directionality. In the main analysis we treat links between enzymes as directed edges, in which an input enzyme catalyzes a reaction producing a metabolite that an output enzyme uses as a substrate. In reality however, the direction of metabolic flux through specific enzymes is often variable, and may be dependent on a number of factors such as metabolite concentration or other environmental conditions. Thus, we constructed an alternative network in which all edges were considered undirected.

Finally, while the use of KEGG has become widespread as a comprehensive and accurate source of metabolic functional information, other databases (e.g. SEED [79]) provide alternative versions of enzyme annotations and their associated reactions. To demonstrate that our results are independent of the KEGG metabolic database altogether, we constructed an alternative network, based exclusively on annotations and reaction data from the SEED annotation system and the MG-RAST analysis server [54]. Specifically, SEED-based enzyme annotations for reads found in the metagenomes of the 124 unrelated individuals were downloaded from MG-RAST (*http://metagenomics.anl.gov/metagenomics.cgi?page=DownloadMetagenome&metagenome=4448044.3*), resulting in 1296 unique enzyme annotations. An odds ratio was calculated for each enzyme based on the pooled count of annotated reads in lean vs. obese samples. Metabolic reactions associated with each enzyme (using their EC number) were downloaded from the modelSEED [159], a tool designed to generate automated genome-scale metabolic models (*http://seed-viewer.theseed.org/ModelSEEDdownload.cgi?biochemCompounds=1*). As MG-RAST does not currently connect directly to the modelSEED framework, we used this modelSEED reaction data to directly generate a metabolic network of 1,005 connected enzymes and projected MG-RAST based

odds ratio scores onto this modelSEED based network. For this analysis, information regarding the cellular location and conformation of various reactions was ignored. Currency metabolites were excluded as described above. Due to the high density of the SEED-based network, an additional set of currency metabolites (NADP, NADPH, CoA, UDP, SAM, AMP, S-Adenosylhomocysteine) was excluded by identifying the top 0.1% of metabolites according to their frequency in all enzyme-associated metabolic reactions.

We find that all four of these alternative networks support our results regarding the relationship between obesity-associated enzymes and centrality (Table S3). Specifically, in each network, we find that obesity-associated differential abundance scores are negatively correlated with enzyme centrality, and that obesity-associated enzymes have lower centrality than other enzymes and are over-represented in the peripheral tier (Table S3).

As further validation of our results, we examined a null version of the network in which edges were randomly rewired (while preserving the in- and out-degree of each node) according to the algorithm outlined in [160]. In contrast to the patterns observed in the real network, we find that of 100 versions of this randomized network, none demonstrated the same low centrality of obesity-associated enzymes, nor the over-representation of obesity-associated enzymes in the peripheral tier. This suggests that the relationship between obesity-associated enzymes and centrality is dependent on the specific network structure dictated by metabolic interactions within the human metagenome.

A.1.10 Seed Set Analysis

The metabolic seeds of the network were determined computationally according to the framework outlined in [75]. This framework uses a novel graph-theory based algorithm to analyze the topology of metabolic networks and to infer the set of compounds that are exogenously acquired. This set (termed ‘seed set’) reflects the metabolic interface between the organism and its surroundings, approximating its environment. Applying this algorithm to the metabolic networks of hundreds of species, this framework was used to generate a large-scale dataset of predicted environments. Seed sets were shown to successfully characterize the biochemical environment of microbial species and to correlate with several basic properties characterizing their environments. Previous studies further used this framework to

identify universal patterns of adaptation of organisms to their niches [75], to predict interactions between microbial species and their hosts [76], and to quantitatively characterize ecological strategies across a large array of microbial species [81]. While the seed algorithm was designed initially to analyze metabolite-based networks, here we use it to determine the seeds in enzyme-based networks. Identified seeds therefore represent enzymes operating on exogenously acquired compounds (rather than the exogenously acquired compounds themselves).

Over- or under-representation of seeds among the various enzyme sets (e.g. obesity-associated enzymes) was determined using a hypergeometric enrichment test. Specifically, for each such set, a p-value was assigned based on the calculated probability of observing *S* or more seeds in a random set of *N* enzymes, given the number of seeds in the entire network. Here, *N* is the total number of enzymes in the set and *S* is the number of enzymes in this set identified as seeds. Over-representation of seeds among the 350 obesity-associated enzymes was further validated by assessing the number of seeds in 10,000 randomly selected sets of 350 enzymes ($p<2x10^{-4}$).

A.1.11 Functional Annotation and Functional Enrichment Analysis

The KEGG BRITE database was mined to annotate each enzyme with zero or more functional categories. A Hypergeometric enrichment test was used to assess the over- or under-representation of each functional category within the set of enriched or depleted enzymes (Table S4). P-values were obtained for each function independently by calculating the probability of observing S or more (for *over*-representation analysis) enzymes annotated with a given function in a random set of N enzymes, given the frequency of that functional annotation in the entire network. Here, S refers to the number of enzymes annotated with this function within a given set of enzymes (e.g., obesity-enriched enzymes) and N refers to the total number of enzymes in this set. Both over- and under-representation of functional annotations were considered.

A.1.12 Single Genome Analysis

The reference genomes of 326 human-associated microbial species were downloaded from the Integrated Microbial Genomes (IMG) Database [161] by searching for species labeled as being part of the

Human Gut Microbiome Initiative or the Human Microbiome Project (Table S5). These genomes were annotated with KEGG orthologous groups and the number of associated genomes was recorded for each enzyme within our network. Correlations were then assessed between enzyme centrality or differential abundance and prevalence within this set of genomes. The reference genome prevalence of the 350 obesity-associated enzymes was further compared to the mean genome prevalence of 10,000 randomly selected sets of 350 enzymes ($p<10^{-4}$).

A.1.13 Topology-Based Biomarker Analysis

In addition to enhancing the digestion of complex carbohydrates and proteins, the gut microbiome encodes a diverse set of genes for xenobiotic metabolism [57] influencing the bioavailability, toxicity, and activity of therapeutic drugs and dietary supplements (e.g [162]). Interestingly, both the obesity- and IBD-associated gene sets with distinct topological features (Tables S2A-D) encode enzymes for xenobiotic metabolism, most notably those for the metabolism of choline and p-cresol (enriched and depleted in IBD/obesity respectively). Recent work has linked microbial metabolism of choline to cardiovascular disease [163], while p-cresol produced by the gut microbiome may interfere with the sulfonation of acetaminophen [164]. Follow-up studies specifically targetted at links between xenobiotic metabolism, obesity, and IBD will be necessary to confirm these associations and to determine if they contribute to disease.

A.1.14 Significance Analysis of Observed Differences in Global Network Properties

In the main text we demonstrated that the network that represents obese samples is less modular than the network that represents lean samples. Here, we confirm that this result is significant and is not expected at random from multiple individual realizations of networks with similar topological properties. To this end, we randomly shuffled the host state labels (e.g., lean vs. obese) 1,000 times, constructed lean- and obese-specific networks based on these shuffled sample labels, and used these networks to generate a null distribution of modularity scores expected for a random set of samples. We compared the modularity scores obtained for the real lean and obese networks to modularity scores obtained with the shuffled sample labels. We find that the difference between the modularity scores of our real lean- and

real obese-specific networks is significantly greater than the pairwise differences of shuffled lean- and obese-specific networks ($p<0.027$; Fig. 2.3*B*). Interestingly, when examining the modularity of lean- and obese-specific networks separately, we find that while the modularity of the real obese-specific network is significantly lower than the modularity of the shuffled-obese networks ($p<0.05$), the modularity of the lean-specific network is higher than most (83%) shuffled-lean networks but is not significantly different (Fig. A.9*A*-*B*). This suggests that the observed difference in the modularity between lean- and obese-specific networks can be attributed mostly to obese samples which deviate from the 'normal' network topology.

We further confirmed the significance of the reported difference in modularity between lean- and obese-specific networks using an alternative null distribution generated by shuffling network edges, as described in [160] and in the section "*Robustness to Alternative Network Construction Methods"* above. For each iteration, network edges in the microbiome-wide network (composed of enzymes found in the union of all samples) were randomly shuffled, and sub-networks were extracted, each representing the set of enzymes found in samples from a specific host state. In total, 1,000 pairs of these shuffled lean- and obese-specific networks were generated. For both obese-specific and lean-healthy-specific networks, the shuffled networks are significantly less modular than the real network. Yet, the observed *difference* in modularity between the true lean- and obese-specific networks is significantly greater than the modularity differences obtained from the shuffled networks ($p<2.0\text{x}10^{-3}$; Fig. A.10).

Other topological features that seem to differ (though not as strongly as the difference in modularity) between host state-specific networks are further illustrated in Fig. A.11*A*-*D*. Specifically, the density of IBD-specific networks appears to be higher than the density of lean-healthy networks, though not significantly so (Figs. A.11*A*, A.11*C*).We also find that the node count of the obese-specific network is significantly lower than the node count of lean-specific networks ($p<0.001$; Fig. A.11*D*).

A.2 Supplementary Figures

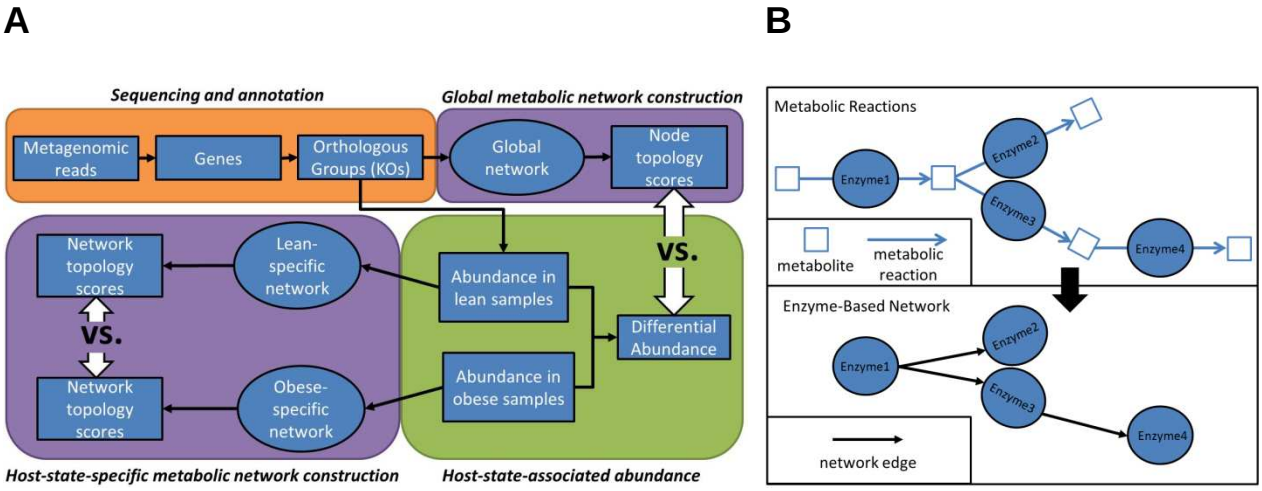


Fig. A.1. Network-based analysis of community metabolism. **(A)** A flowchart of the analysis presented in this study. Briefly, metagenomic reads from various samples are mapped to orthologous groups (KOs). The differential abundance score of each enzyme (KO) is calculated by comparing the enzyme abundance in samples from different host-states. A global metabolic network is constructed from the entire sample-wide set of enzymes, as depicted in **(B)**. Each enzyme is assigned a set of substrate and product metabolites (white squares) according to the reactions catalyzed by the enzyme as annotated in KEGG. A search is performed to identify all enzyme pairs in which a product metabolite of one enzyme is a substrate metabolite of the other. Directed network edges are then drawn between each identified enzyme pair. Once the network has been generation, topological features of individual enzyme nodes are compared to the enzyme's differential abundance. A parallel analysis involves the construction of a separate network from enzymes found in the subset of samples from a specific host state. Network-wide topological features are compared across these host-state-specific networks.

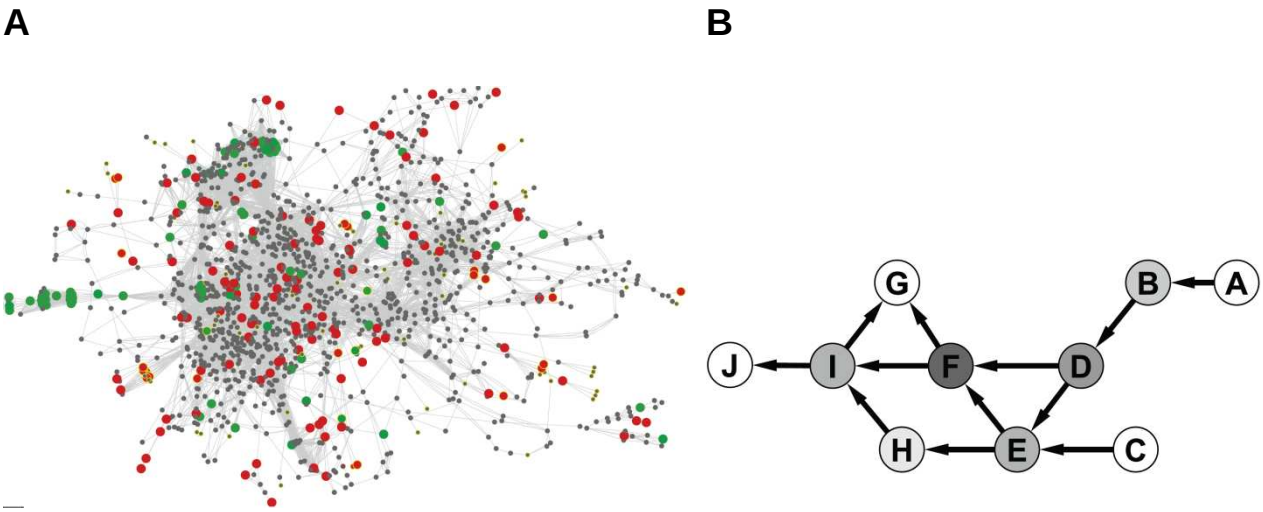


Fig. A.2. A community-level metabolic network. Nodes represent enzymes and edges connect enzymes that catalyze successive metabolic steps. A full network of the enzymes (n=1563) found in a set of 124 human gut microbiomes is shown in **(A)**. Enzymes that are associated with obesity appear as larger colored nodes (red=enriched, green=depleted). Nodes outlined in yellow were identified as network seeds. A simple synthetic network is illustrated in **(B)** to highlight the various topological properties examined. Betweenness centrality quantifies the topological importance of a node, while the clustering coefficient measures the tendency of the node's immediate neighbors to form a fully connected sub-network. The indegree and outdegree of a node denote the number of edges terminating or originating in a node, respectively. In this simple network, the shading of the node corresponds to its centrality. Here, node B has a centrality score of 0.194, in-degree 1, out-degree 1, and clustering coefficient 0. In contrast, node F has a centrality score of 0.361, indegree 2, outdegree 2, and clustering coefficient 0.17. Networks were visualized using the Cytoscape software [105].

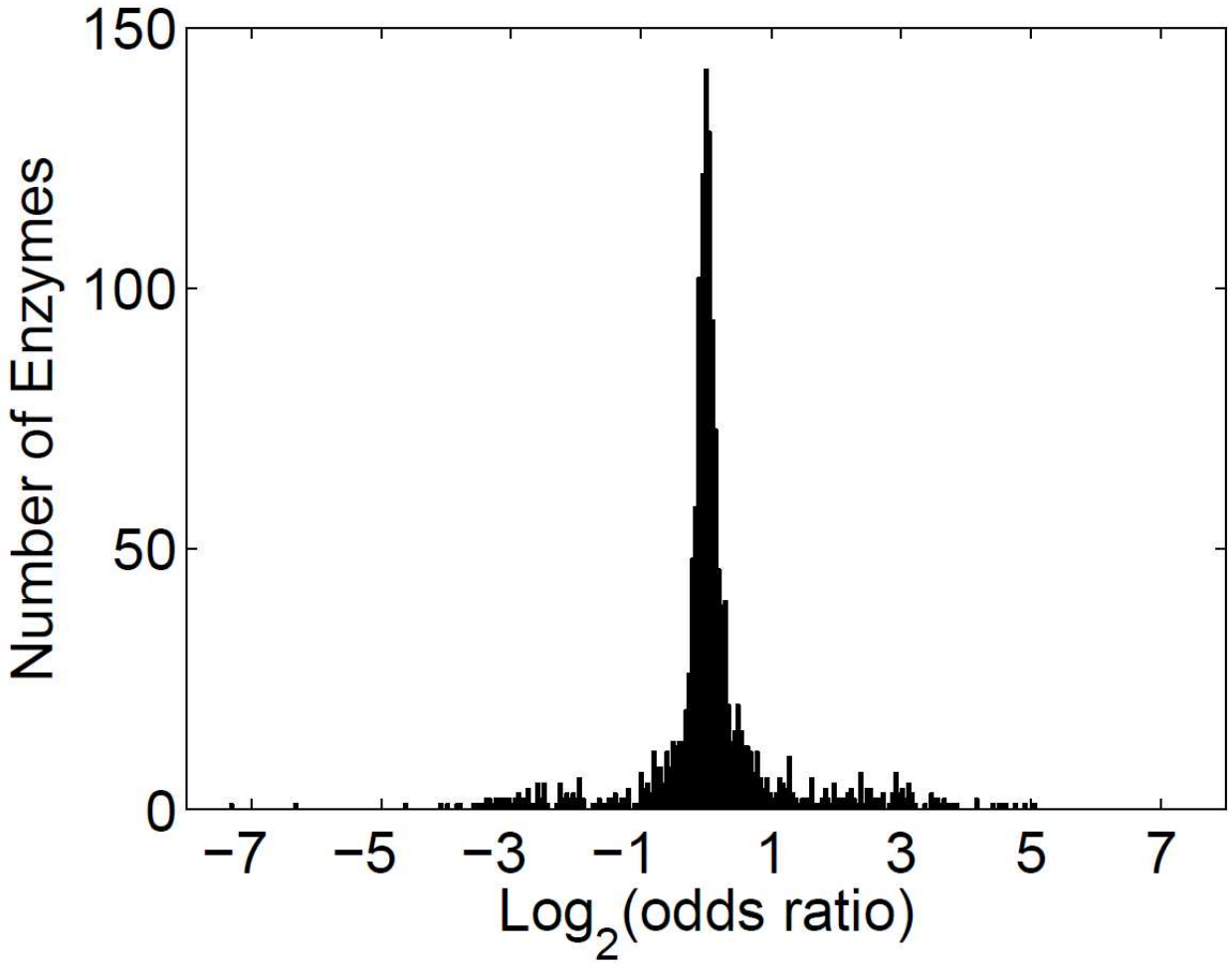


Fig. A.3. Histogram of differential abundance across the 1563 enzymes found in the global metabolic network. Differential abundance is defined as the $\log_2$ of the odds ratio of pooled enzyme abundance in obese vs. lean samples. Enzymes with a fold change greater than 1 were classified as 'enriched.' Enzymes with a fold change less than -1 were classified as 'depleted.'

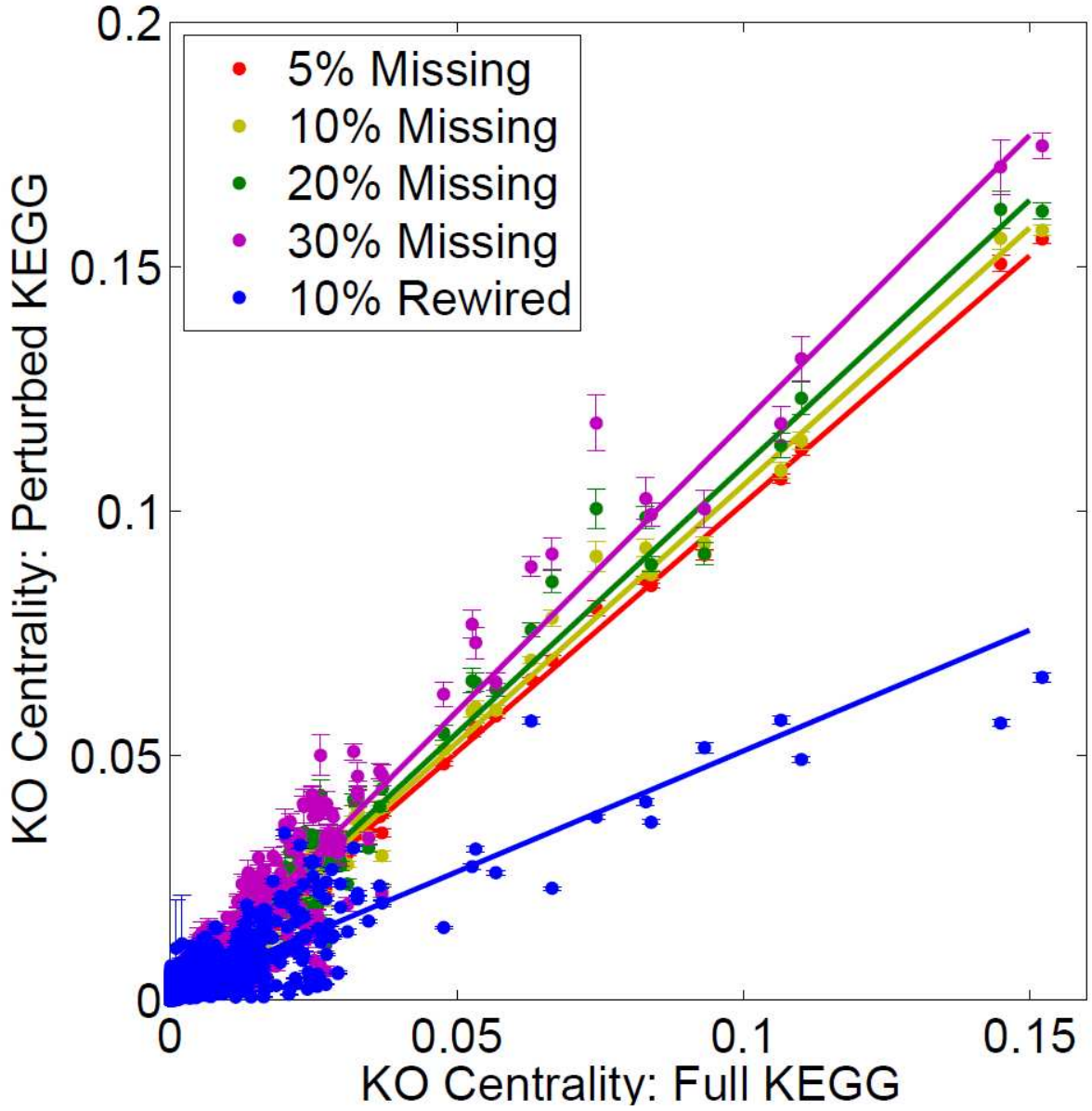


Fig. A.4. The effect of missing or erroneous annotation data on centrality values in the metabolic network. To determine the effect of incomplete annotation data, we deleted a randomly selected subset of 5%, 10%, 20%, or 30% of the annotations in KEGG, and re-constructed global metabolic networks from the remaining annotated enzymes. Similarly, to simulate the effect of erroneous annotations, we rewired a random subset of 10% of the edges in the network. The mean centrality score and standard deviation (over 100 iterations) of each KO in these perturbed networks is plotted against its centrality score in the fully-annotated network, demonstrating a relatively small sensitivity of centrality scores to annotation inaccuracies. In each case, KO centrality scores in the perturb network are strongly correlated with the original centrality scores. Spearman correlation coefficient: 0.998 (5% deleted), 0.996 (10% deleted), 0.984 (20% deleted), 0.969 (30% deleted), 0.788 (10% rewired).

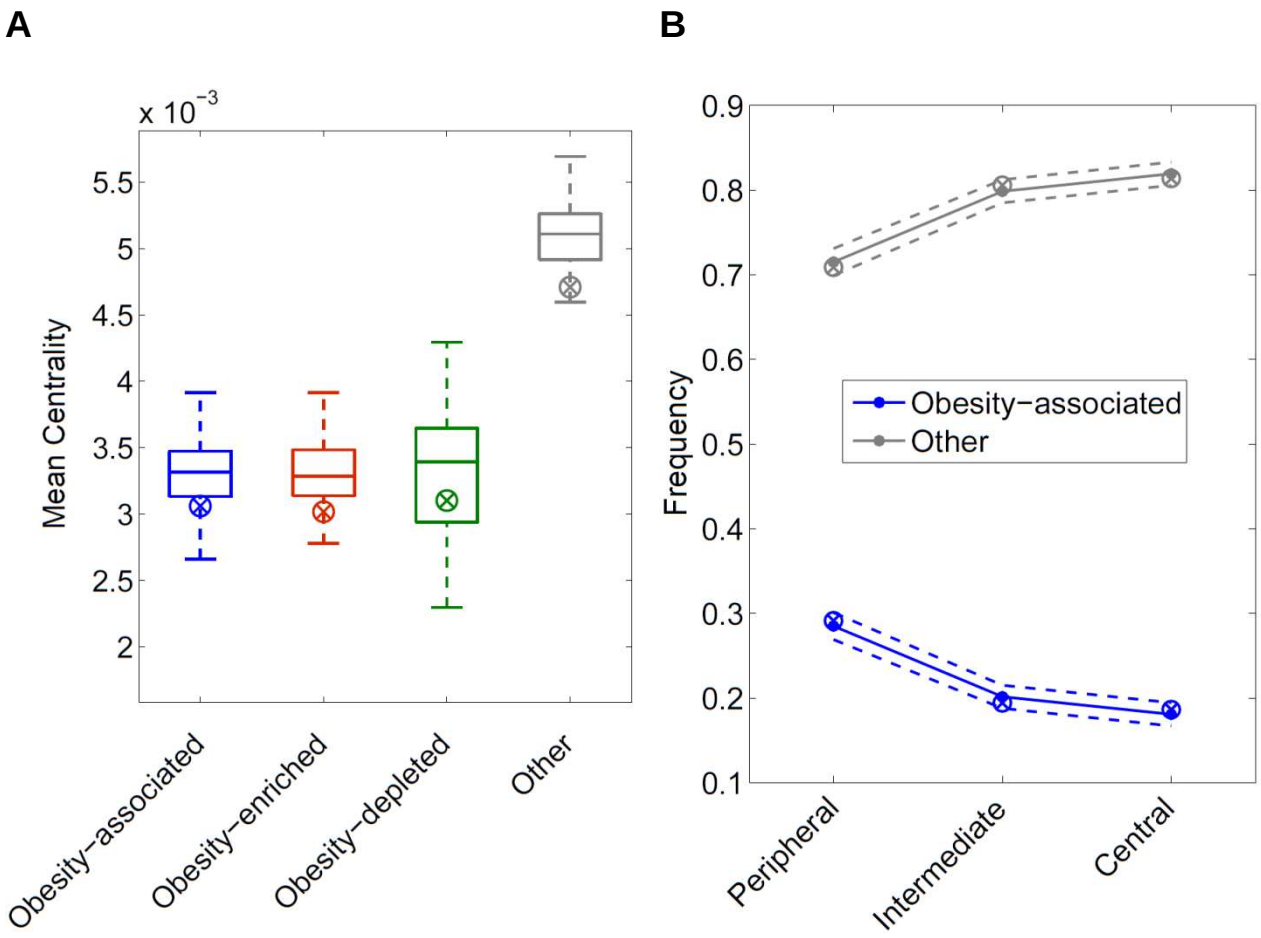


Fig. A.5. The effect of incomplete enzyme annotation on the relationship between centrality and obesity-associated enzymes. **(A)** Boxplots depicting the distribution of the mean centrality score of enzymes in 100 iterations of networks constructed after randomly deleting 10% of the annotations in KEGG. Mean centrality is calculated across enzymes grouped according to their association with obesity. The mean centrality of these enzyme groups in the fully-annotated network is plotted as a circled X's. Evidently, in these perturbed networks, the centrality of obesity-associated enzymes (as well as obesity-enriched and obesity-depleted enzymes separately) is still significantly lower than the centrality of other enzymes. **(B)** Frequency of obesity-associated vs. other enzymes in different centrality tiers (peripheral, intermediate, and central) in perturbed networks constructed after randomly deleting 10% of the annotations in KEGG. Plotted above are the mean (filled circles) and standard deviation (dotted lines) over 100 such perturbed networks. Circled X's represent the frequency in the fully-annotated network.

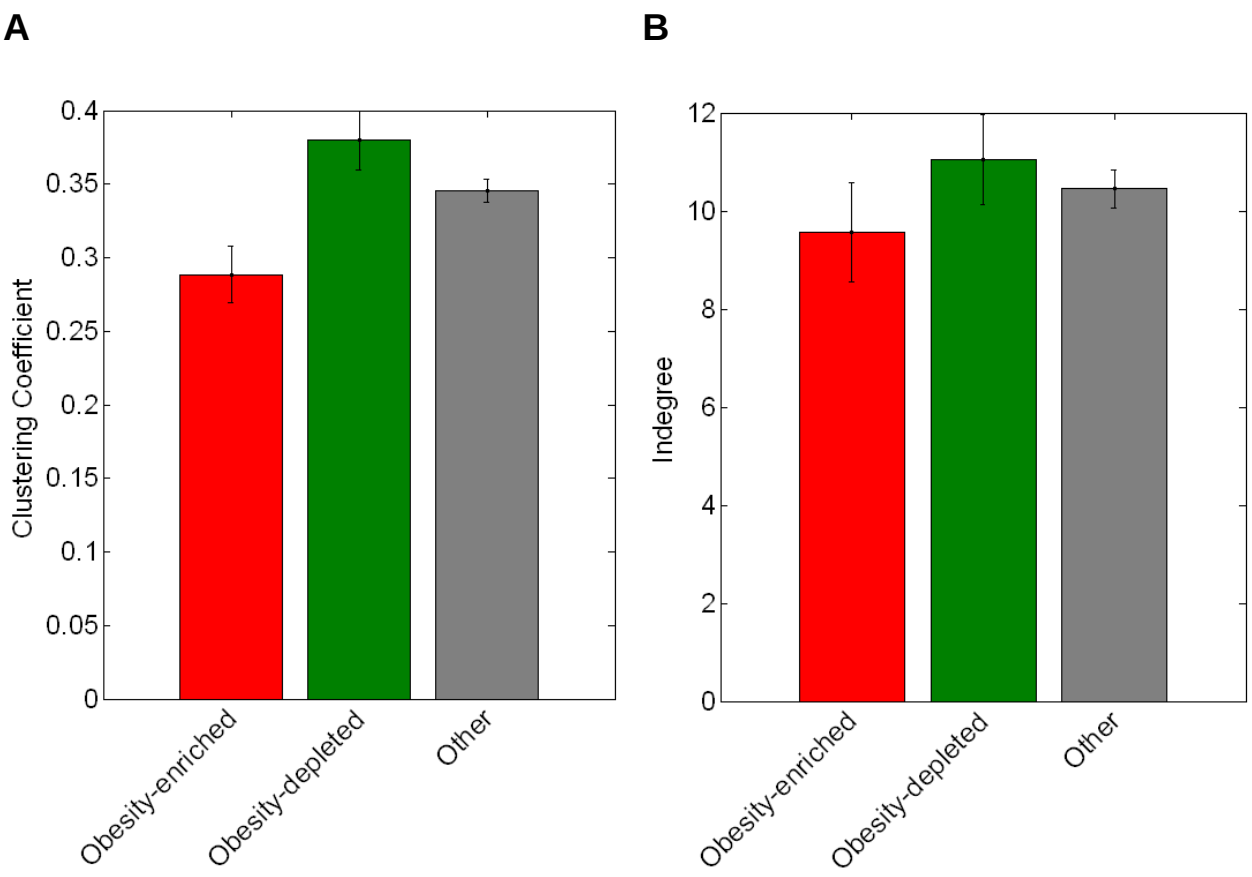


Fig. A.6. Mean and standard error of the **(A)** clustering coefficient and **(B)** indegree of enriched, depleted, and other enzymes in IBD microbiomes. As with obesity-associated enzymes, IBD-enriched enzymes tend to have a lower clustering coefficient ($p<0.015$) and lower indegree compared to other enzymes, while IBD-depleted enzymes tend to have a higher clustering coefficient and higher indegree. These differences mostly fall short of statistical significance, possibly due to smaller sample size.

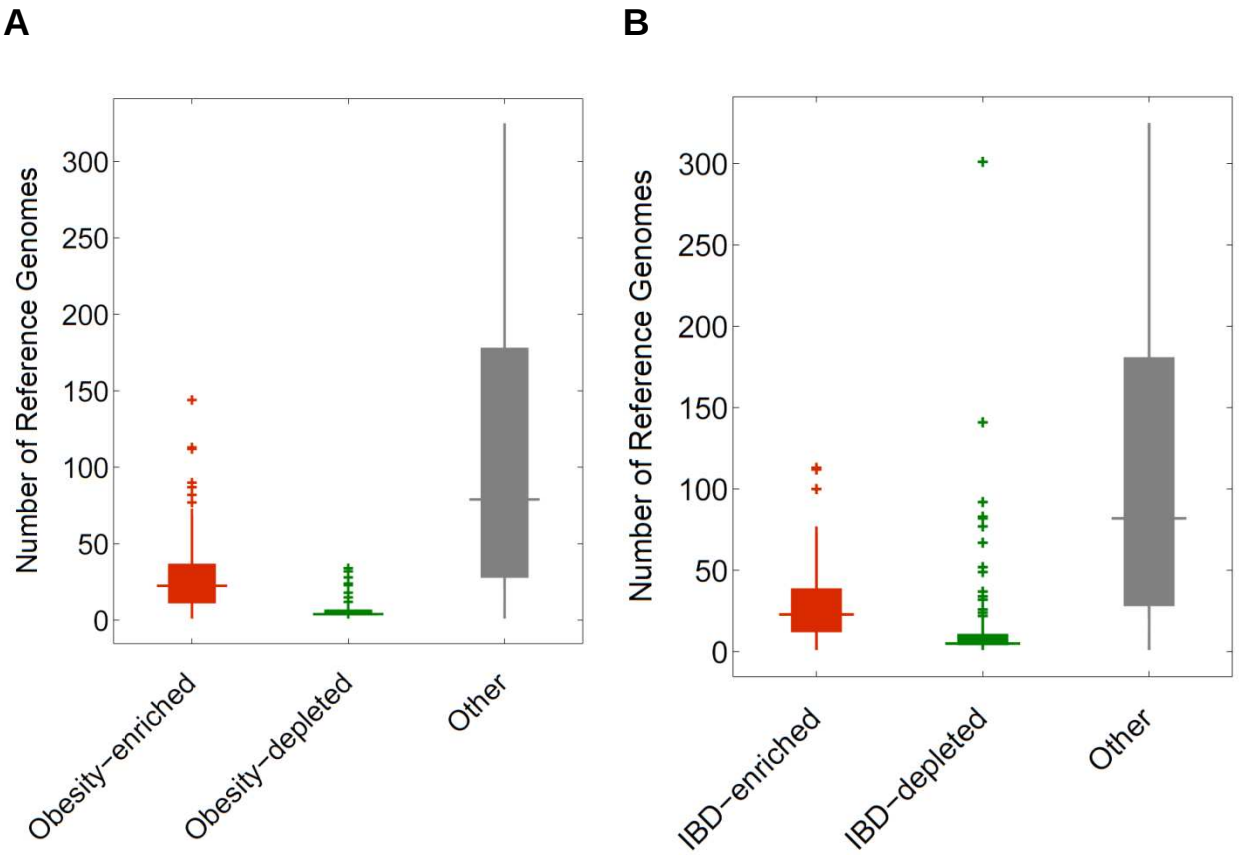


Fig. A.7. Boxplots of the number of reference genomes (n=326) associated with enzymes of different classes in **(A)** obese and **(B)** IBD microbiomes. Enzymes associated with either obesity or IBD are present in far fewer reference genomes ($p<2.0x10^{-55}$ [obese]; $p<2.410^{-57}$ [IBD]; Wilcoxon rank-sum test).

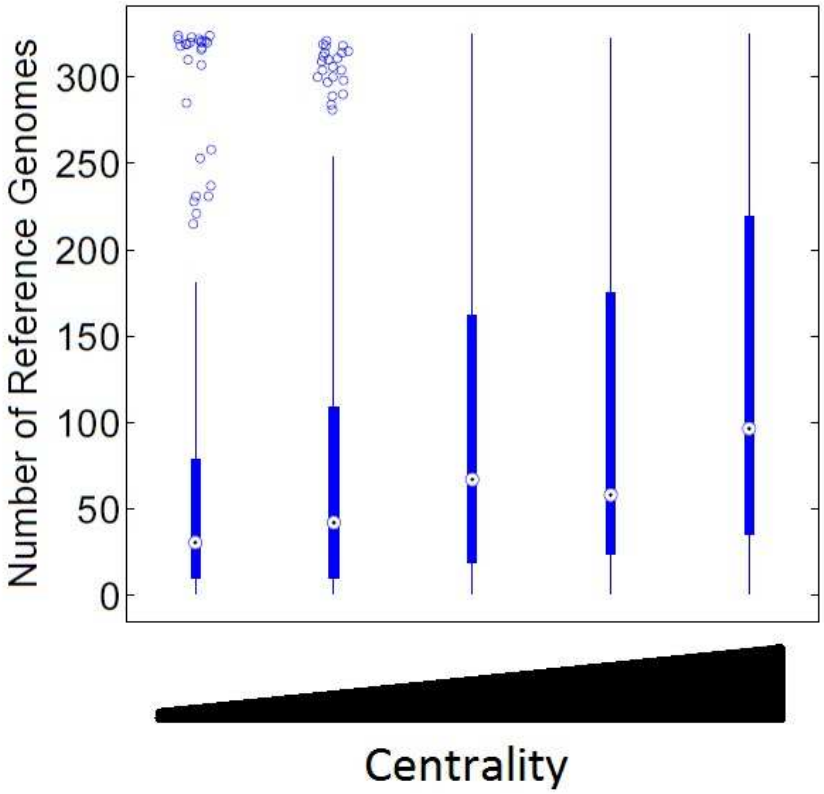


Fig. A.8. Boxplots of the number of reference genomes (n=326) associated with enzymes in different centrality tiers. Boxes are arranged in order of increasing centrality from left to right. Centrality is positively correlated with the number of associated reference genomes ($R = 0.23$, $p<6.0x10^{-18}$; Spearman correlation test).

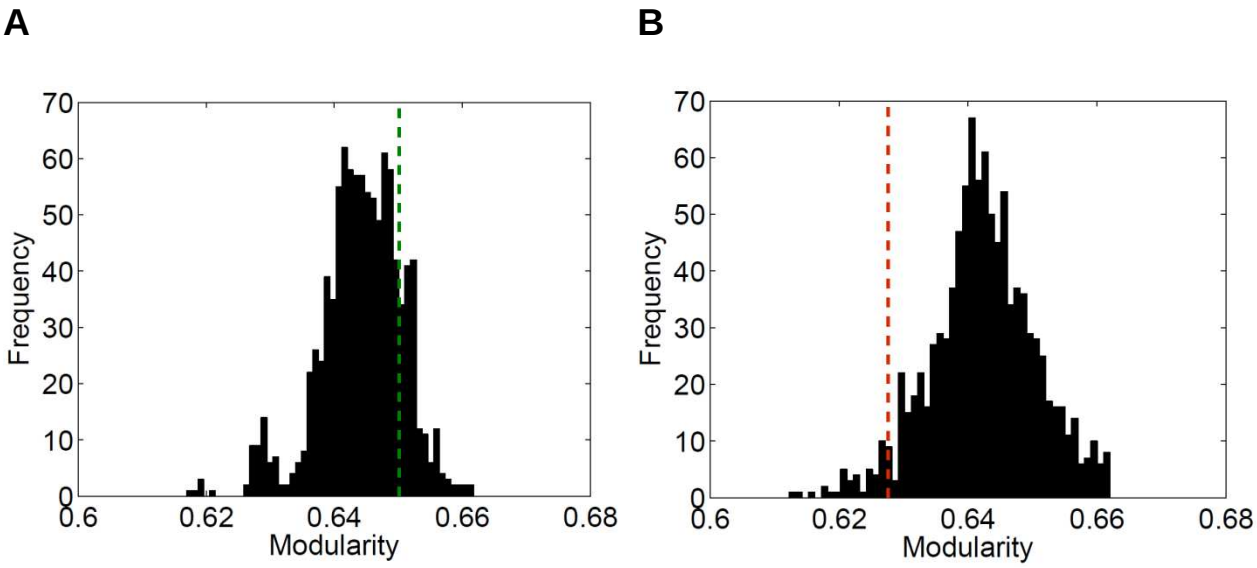


Fig. A.9. Modularity of **(A)** lean/healthy-specific and **(B)** obese-specific metabolic networks compared to null distributions. To create null distributions, sample labels were randomly shuffled 1,000 times. For each shuffled set, separate lean/healthy and obese-specific networks were constructed and the modularity of each network was calculated (see Supporting Text for more information). Plotted above are histograms of the modularity of these shuffled networks. The modularity of the true lean/healthy-specific network (green dotted line) is higher than the modularity of shuffled networks in most cases (83%) but is not significantly different. In contrast, modularity of the true obese-specific network (red dotted line) is significantly lower than expected ($p<0.05$).

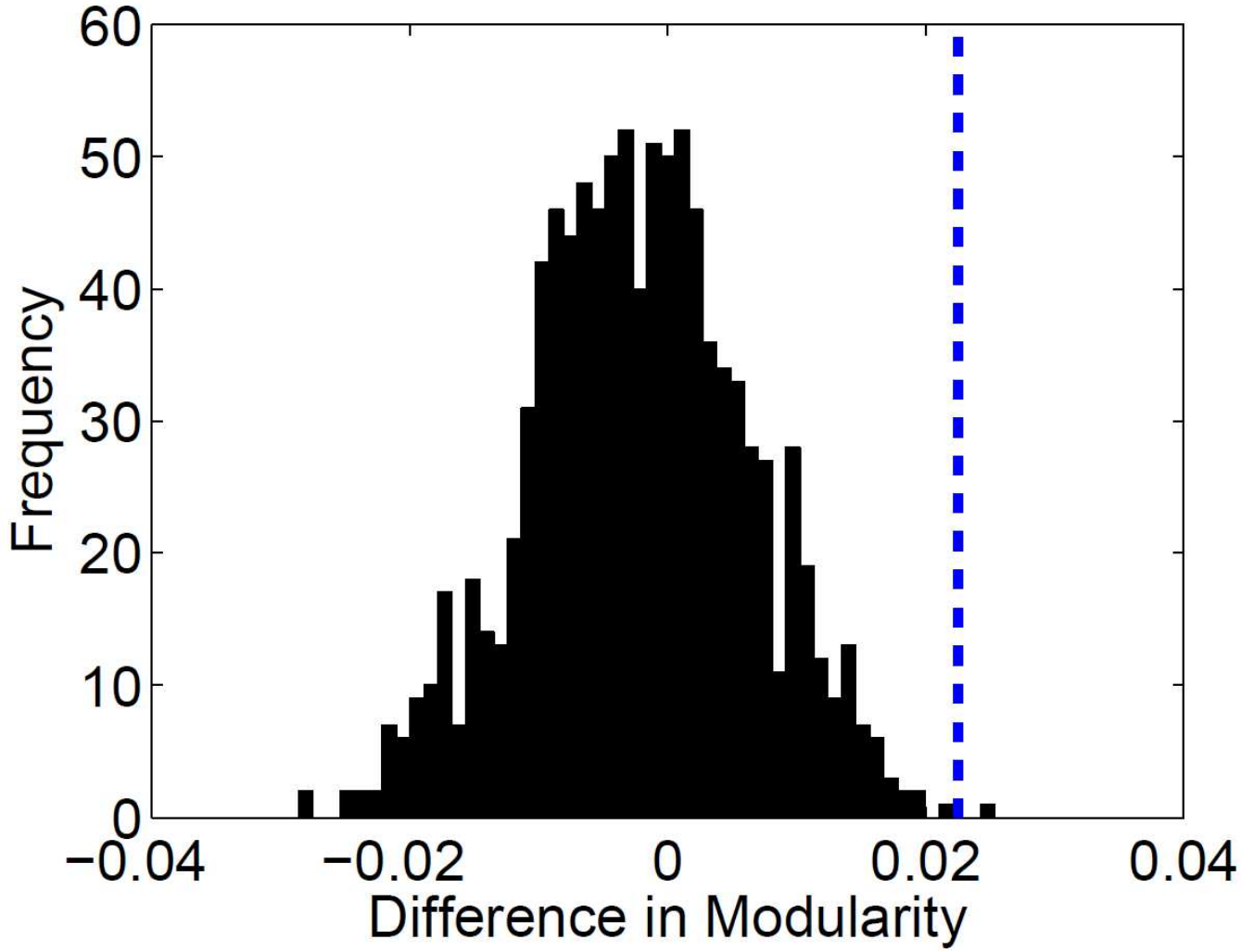


Fig. A.10. The difference between the modularity of the obese-specific and lean-healthy-specific network is plotted (dashed blue line) against a null distribution of differences from shuffled networks. To create this null distribution, we created 1,000 pairs of shuffled obese-specific and shuffled lean-healthy-specific networks by randomly rewiring edges in the microbiome-wide network while preserving the indegree and outdegree of each node (see SI Text), and extracting sub-networks corresponding to the enzymes found in samples from a specific host state. The difference in modularity between each corresponding pair of shuffled obese-specific and shuffled lean-healthy-specific network was calculated. In both obese-specific and lean-healthy-specific networks, shuffled networks are significantly less modular than the real network. Yet, the observed difference in modularity in the true networks is significantly greater than the expected difference in the shuffled networks ($p<2.0 \times 10^{-3}$).

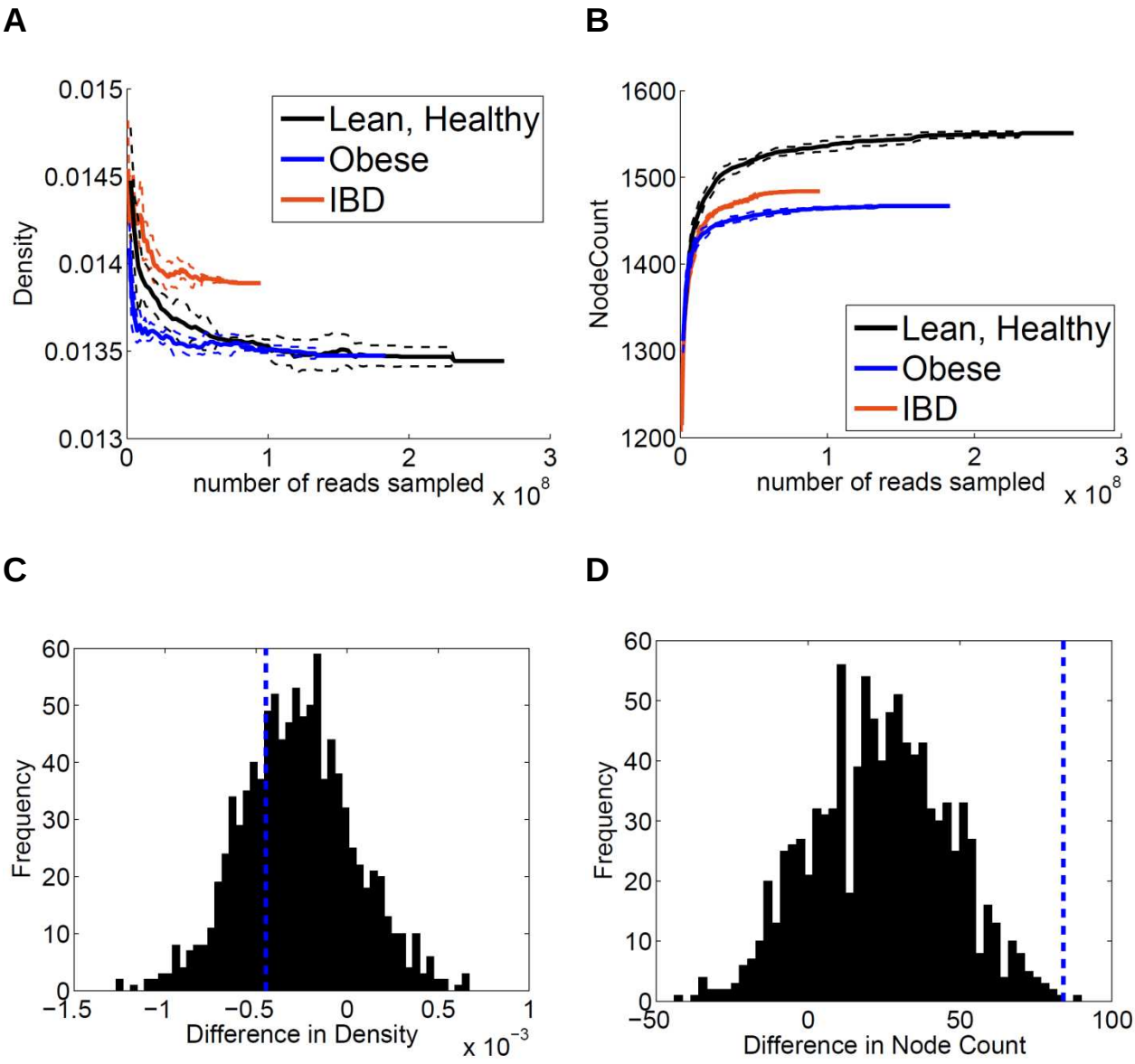


Fig. A.11. Rarefaction analysis of network density **(A)** and node count **(B)** for host-state-specific networks. The **(C)** difference in density between the lean/healthy-specific and IBD-specific networks and **(D)** difference in node count between the lean/healthy-specific and obese-specific networks is shown (dotted line) compared to null distributions. Null distributions were generated by shuffling sample labels 1,000 times (see A.1 Supplementary Text and Fig. A.9 legend). The observed difference between the node count of the lean/healthy and obese-specific networks is statistically greater than expected compared to this null distribution (p=0.001).

A.3 Supplemental Tables

Table A.1. Validation of main results

ANALYSIS	ASSAY 1*	ASSAY 2**	ASSAY 3***
Main Results			
Obesity	1.34E-11	8.93E-06	5.55E-06
IBD	1.94E-09	9.58E-06	4.74E-05
Alternative Dataset			
Twins[#] dataset (Obesity)	9.7E-08	1.6E-04	1.2E-03
Non-Transport Enzymes Analysis			
Omitting transport enzymes (Obesity)	6.59E-10	1.58E-05	3.05E-05
Omitting transport enzymes (IBD)	2.79E-08	3.13E-05	1.61E-04
Robustness to Population Structure			
Danish samples only (Obesity)	2.81E-09	2.50E-05	5.50E-05
Spanish samples only (IBD)	8.94E-10	8.05E-06	1.36E-06
Alternative Enrichment Metrics			
Ranksum (Obesity)	N/A	1.53E-02	1.00E-02
Presence/absence hypergeometric (Obesity)	N/A	3.57E-03	2.18E-03
Difference in rank (Obesity)	N/A	1.28E-03	1.72E-02
Jensen–Shannon divergence (Obesity)	N/A	1.76E-06	1.75E-04
Robustness to Noise in Read Count Data			
Read Threshold = 5 (Obesity)	2.26E-12	6.65E-07	2.46E-07
Read Threshold = 25 (Obesity)	7.47E-09	1.23E-05	8.83E-06
Read Threshold = 50 (Obesity)	3.46E-08	1.30E-06	4.90E-06
Consistently Enriched/Depleted (Obesity)	N/A	2.83E-02	2.44E-02
Read Threshold = 5 (IBD)	2.03E-09	1.20E-05	1.26E-05
Read Threshold = 25 (IBD)	1.18E-06	7.11E-06	6.00E-05
Read Threshold = 50 (IBD)	1.06E-05	1.23E-05	7.68E-05
Consistently Enriched/Depleted (IBD)	N/A	3.09E-02	1.64E-02
Alternative Network Construction Models			
All KEGG Network (Obesity)	4.26E-06	1.25E-03	4.60E-06
Undirected Edges (Obesity)	2.53E-06	1.12E-03	2.66E-03
Omitting Currency Metabolites (Obesity)	2.07E-09	9.62E-04	8.35E-04
SEED-based Network (Obesity)	6.04E-03	3.48E-02	1.13E-02
All KEGG Network (IBD)	1.35E-04	6.80E-03	1.95E-03
Undirected Edges (IBD)	7.67E-05	8.59E-03	7.69E-03
Omitting Currency Metabolites (IBD)	2.22E-07	7.67E-05	1.70E-03
SEED-based Network (IBD)	1.30E-03	1.10E-03	1.89E-03

Three assays were used to validate the association between enzyme centrality and differential abundance.
*Assay 1: p-value of the Spearman correlation between centrality and differential abundance score
**Assay 2: p-value of the Wilcoxon ranksum test comparing the centrality of host-state associated enzymes to the centrality of other enzymes
***Assay 3: p-value of the Hypergeometric enrichment test measuring over-representation of host-state associated enzymes in the peripheral tier of the network
[#]Turnbaugh PJ et al. (2009) A core gut microbiome in obese and lean twins. Nature 457:480-4.

Table A.2. Functional enrichment of host-state associated enzymes according to KEGG BRITE classes

Centrality Class	Association Class	Over-Represented Functions	Under-Represented Functions
All Enzymes	Obese-Depleted	Cell Communication** Replication and Repair** Transcription*** Neurodegenerative Diseases*** Metabolism of Cofactors and Vitamins* Nucleotide Metabolism** Circulatory System* Immune System**	Membrane Transport** Amino Acid Metabolism * Carbohydrate Metabolism *** Biosynthesis of Polyketides and Terpenoids* Energy Metabolism*
	Obese- Enriched	Membrane Transport*	Transcription* Glycan Biosynthesis and Metabolism*
	IBD-Depleted	Cell Communication** Transcription*** Neurodegenerative Diseases** Nucleotide Metabolism* Circulatory System* Immune System*	Membrane Transport** Carbohydrate Metabolism *** Energy Metabolism**
	IBD-Enriched	Membrane Transport*	Metabolism of Cofactors and Vitamins* Nucleotide Metabolism* Replication and Repair* Transcription*
Peripheral Enzymes	Obese-Depleted	Transcription*** Nucleotide Metabolism*** Signal Transduction* Replication and Repair*** Neurodegenerative Diseases** Immune System *	Carbohydrate Metabolism*** Amino Acid Metabolism** Membrane Transport* Energy Metabolism*
	Obese- Enriched	Membrane Transport***	Metabolism of Cofactors and Vitamins*
	IBD-Depleted	Replication and Repair** Transcription*** Neurodegenerative Diseases* Nucleotide Metabolism***	Amino Acid Metabolism** Carbohydrate Metabolism* Energy Metabolism** Metabolism of Cofactors and Vitamins*
	IBD-Enriched	Membrane Transport***	
	All Enzymes	Membrane Transport*** Replication and Repair*** Translation*** Transcription*** Nucleotide Metabolism**	Amino Acid Metabolism*** Carbohydrate Metabolism*** Energy Metabolism* Endocrine System* Metabolism of Cofactors and Vitamins*

Appendix B – Supplementary material for chapter 3

B.1 Supplementary Figures

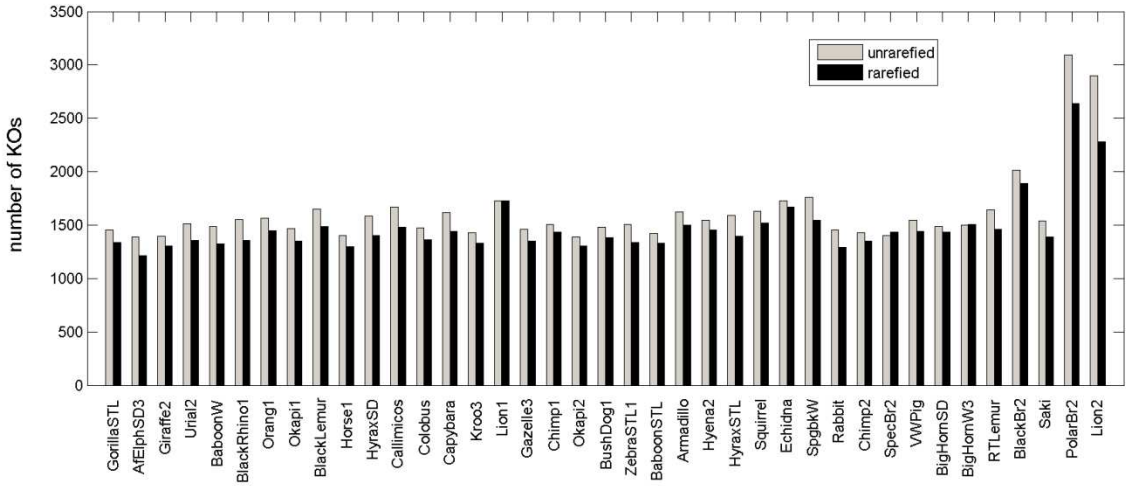


Fig. B.1 Number of KOs detected in mammalian microbiomes. KOs were considered present at a normalized abundance >0.1. Total number of present KOs was assessed before and after rarefaction to 14,236 reads per sample. Samples are sorted by average read length, from shortest (GorillaSTL; 153bp) to longest (Lion2; 413bp).

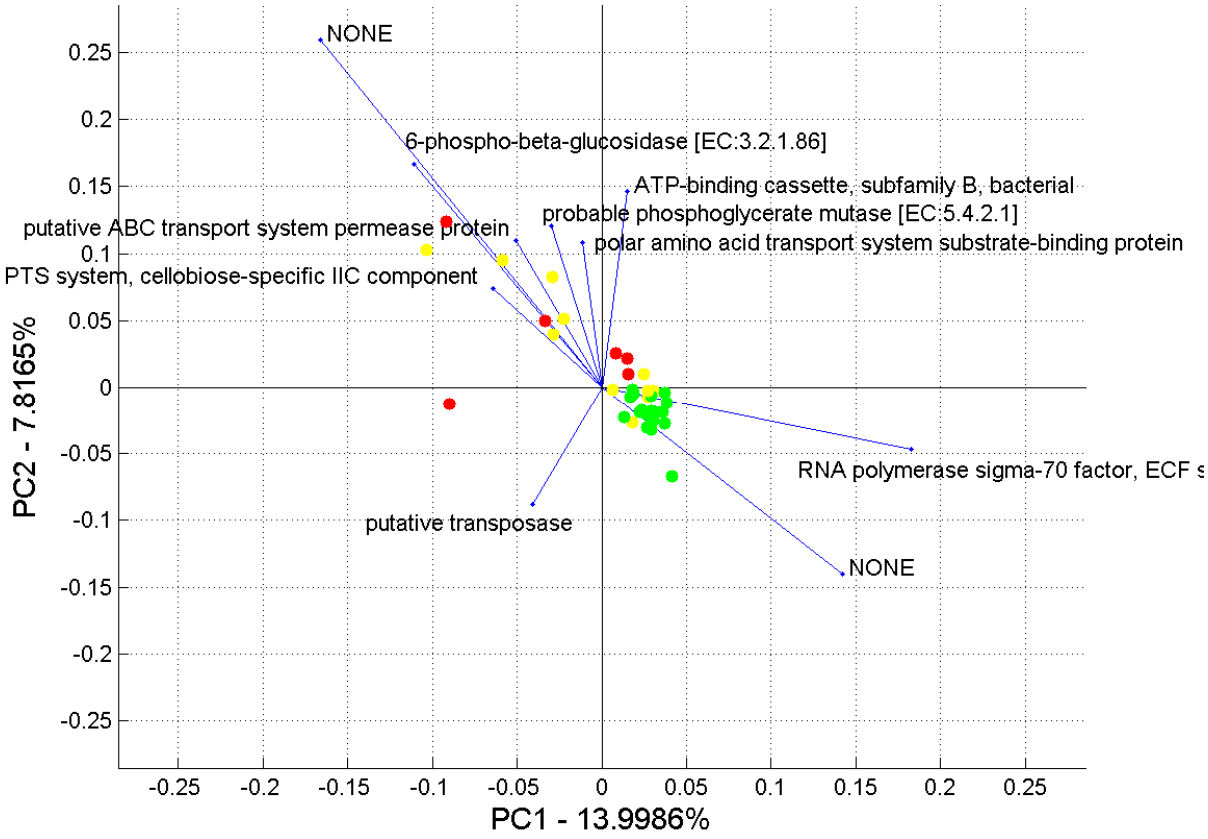


Fig. B.2 Principal component analysis of mammalian microbiome samples based on rarefied KO abundances. Samples are filled circles colored according to diet (green: herbivore, yellow: omnivore, red: carnivore). The 10 KOs with the highest loadings are shown as labeled lines in the biplot.

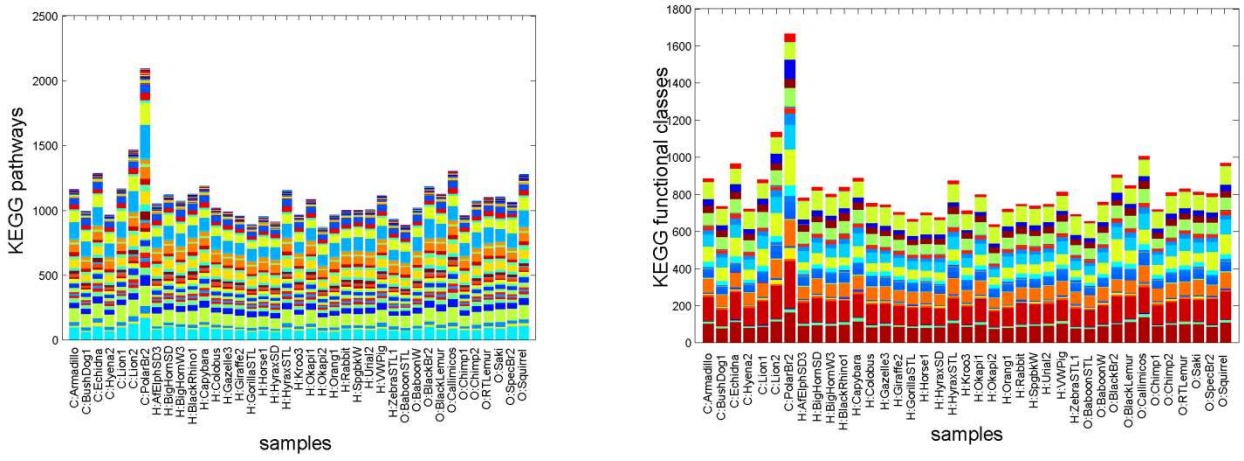


Fig. B.3 Relative abundance of KEGG pathways and functional classes across mammalian microbiomes. Samples are sorted according to diet group (C:carnivore, H:herbivore, O:omnivore).

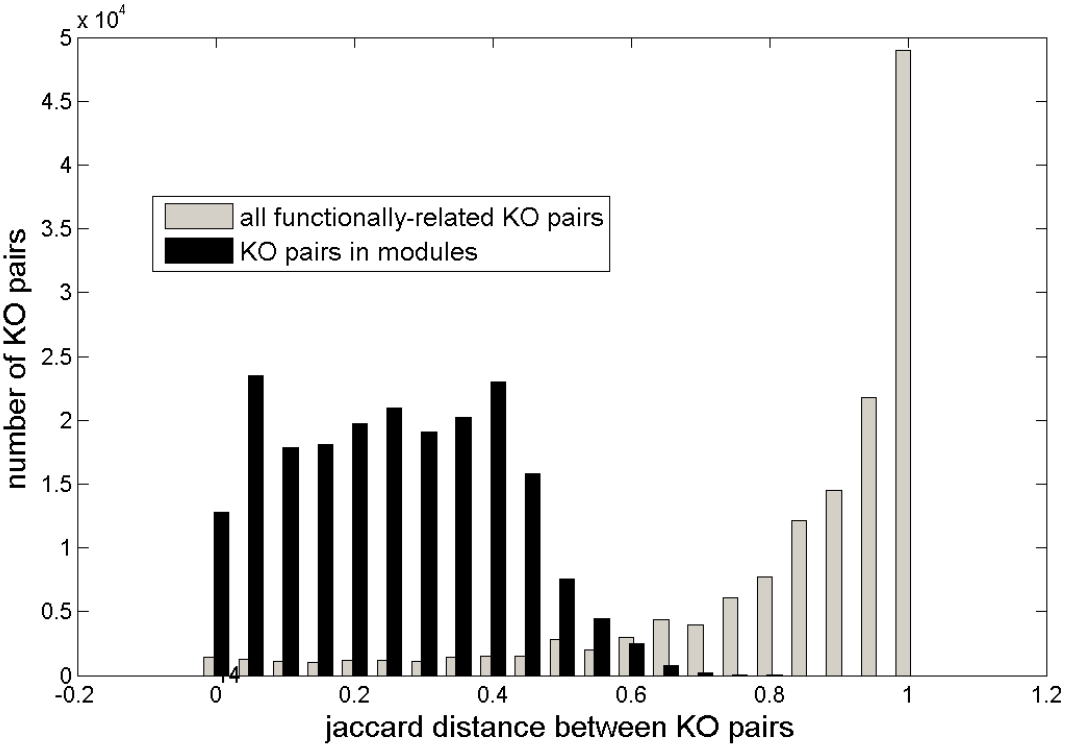


Fig. B.4 Histograms of pairwise KO jaccard distances. The jaccard distance between the presence/absence profiles of each pair of KOs across 39 mammalian hosts was calculated. Shown are distributions for pairs of KOs within the same module (black), and all pairs of functionally related KOs (gray). The two distributions are significantly different according to a Student's t-test at $p<10^{-52}$.

B.2 Supplementary Tables

Table B.1 KOs comprising shared and organism-specific modules.

Module*	KO	KO name	KO pathway(s)
4	K02796	PTS system, mannose-specific IID component	Fructose and mannose metabolism|Amino sugar and nucleotide sugar metabolism|Phosphotransferase system (PTS)|
4	K02809	PTS system, sucrose-specific IIB component [EC:2.7.1.69]	Starch and sucrose metabolism|Phosphotransferase system (PTS)|
4	K02810	PTS system, sucrose-specific IIC component	Starch and sucrose metabolism|Phosphotransferase system (PTS)|
4	K02808	PTS system, sucrose-specific IIA component [EC:2.7.1.69]	Starch and sucrose metabolism|Phosphotransferase system (PTS)|
4	K02795	PTS system, mannose-specific IIC component	Fructose and mannose metabolism|Amino sugar and nucleotide sugar metabolism|Phosphotransferase system (PTS)|
4	K02793	PTS system, mannose-specific IIA component [EC:2.7.1.69]	Fructose and mannose metabolism|Amino sugar and nucleotide sugar metabolism|Phosphotransferase system (PTS)|
4	K02755	PTS system, beta-glucosides-specific IIA component [EC:2.7.1.69]	Phosphotransferase system (PTS)|

4	K02757	PTS system, beta-glucosides-specific IIC component	Phosphotransferase system (PTS)|
4	K02756	PTS system, beta-glucosides-specific IIB component [EC:2.7.1.69]	Phosphotransferase system (PTS)|
4	K02770	PTS system, fructose-specific IIC component	Fructose and mannose metabolism|Phosphotransferase system (PTS)|
4	K02769	PTS system, fructose-specific IIB component [EC:2.7.1.69]	Fructose and mannose metabolism|Phosphotransferase system (PTS)|
4	K02768	PTS system, fructose-specific IIA component [EC:2.7.1.69]	Fructose and mannose metabolism|Phosphotransferase system (PTS)|
5	K02821	PTS system, ascorbate-specific IIA component [EC:2.7.1.69]	Ascorbate and aldarate metabolism|Phosphotransferase system (PTS)|
5	K02779	PTS system, glucose-specific IIC component	Glycolysis / Gluconeogenesis|Amino sugar and nucleotide sugar metabolism|Phosphotransferase system (PTS)|
5	K02778	PTS system, glucose-specific IIB component [EC:2.7.1.69]	Glycolysis / Gluconeogenesis|Amino sugar and nucleotide sugar metabolism|Phosphotransferase system (PTS)|
5	K02818	PTS system, trehalose-specific IIB component [EC:2.7.1.69]	Starch and sucrose metabolism|Phosphotransferase system (PTS)|
5	K02819	PTS system, trehalose-specific IIC component	Starch and sucrose metabolism|Phosphotransferase system (PTS)|
5	K00694	cellulose synthase (UDP-forming) [EC:2.4.1.12]	Starch and sucrose metabolism|
6	K02787	PTS system, lactose-specific IIB component [EC:2.7.1.69]	Galactose metabolism|Phosphotransferase system (PTS)|
6	K02788	PTS system, lactose-specific IIC component	Galactose metabolism|Phosphotransferase system (PTS)|
6	K02760	PTS system, cellobiose-specific IIB component [EC:2.7.1.69]	Phosphotransferase system (PTS)|
6	K02759	PTS system, cellobiose-specific IIA component [EC:2.7.1.69]	Phosphotransferase system (PTS)|
6	K02749	PTS system, arbutin-like IIB component [EC:2.7.1.69]	Glycolysis / Gluconeogenesis|Phosphotransferase system (PTS)|
6	K02750	PTS system, arbutin-like IIC component	Glycolysis / Gluconeogenesis|Phosphotransferase system (PTS)|
6	K02786	PTS system, lactose-specific IIA component [EC:2.7.1.69]	Galactose metabolism|Phosphotransferase system (PTS)|
6	K01220	6-phospho-beta-galactosidase [EC:3.2.1.85]	Galactose metabolism|
7	K00687	penicillin-binding protein 2B [EC:2.3.2.-]	Peptidoglycan biosynthesis|
7	K12554	alanine adding enzyme [EC:2.3.2.-]	Peptidoglycan biosynthesis|
7	K03693	penicillin-binding protein	Peptidoglycan biosynthesis|
7	K05363	serine/alanine adding enzyme [EC:2.3.2.10]	Peptidoglycan biosynthesis|
8	K11618	two-component system, NarL family, response regulator LiaR	Two-component system|
8	K11617	two-component system, NarL family, sensor histidine kinase LiaS [EC:2.7.13.3]	Two-component system|
8	K14982	two-component system, OmpR family, sensor histidine kinase CiaH [EC:2.7.13.3]	Two-component system|
8	K14983	two-component system, OmpR family, response regulator CiaR	Two-component system|
8	K12296	competence protein ComX	Two-component system|
9	K08680	2-succinyl-6-hydroxy-2,4-cyclohexadiene-1-carboxylate synthase [EC:4.2.99.20]	Ubiquinone and other terpenoid-quinone biosynthesis|
9	K03185	2-octaprenyl-6-methoxyphenol hydroxylase [EC:1.14.13.-]	Ubiquinone and other terpenoid-quinone biosynthesis|
9	K02552	menaquinone-specific isochorismate synthase [EC:5.4.4.2]	Ubiquinone and other terpenoid-quinone biosynthesis|Biosynthesis of siderophore group nonribosomal peptides|
9	K02364	enterobactin synthetase component F [EC:2.7.7.-]	Biosynthesis of siderophore group nonribosomal peptides|
9	K01252	bifunctional isochorismate lyase / aryl carrier protein [EC:3.3.2.1]	Biosynthesis of siderophore group nonribosomal peptides|
10	K12525	bifunctional aspartokinase / homoserine dehydrogenase 2 [EC:2.7.2.4 1.1.1.3]	Glycine, serine and threonine metabolism|Cysteine and methionine metabolism|Lysine biosynthesis|
10	K00998	CDP-diacylglycerol---serine O-phosphatidyltransferase [EC:2.7.8.8]	Glycine, serine and threonine metabolism|Glycerophospholipid metabolism|
10	K01521	CDP-diacylglycerol pyrophosphatase [EC:3.6.1.26]	Glycerophospholipid metabolism|
10	K06132	putative cardiolipin synthase [EC:2.7.8.-]	Glycerophospholipid metabolism|
10	K01058	phospholipase A1 [EC:3.1.1.32 3.1.1.4]	Glycerophospholipid metabolism|Ether lipid metabolism|Arachidonic acid metabolism|Linoleic acid

			metabolism|alpha-Linolenic acid ... <Preview truncated at 128 characters>
10	K00631	glycerol-3-phosphate O-acyltransferase [EC:2.3.1.15]	Glycerolipid metabolism|Glycerophospholipid metabolism|
11	K02849	heptosyltransferase III [EC:2.4.-.-]	Lipopolysaccharide biosynthesis|
11	K02844	UDP-glucose:(heptosyl)LPS alpha-1,3-glucosyltransferase [EC:2.4.1.-]	Lipopolysaccharide biosynthesis|
11	K02560	lipid A biosynthesis (KDO)2-(lauroyl)-lipid iva acyltransferase [EC:2.3.1.-]	Lipopolysaccharide biosynthesis|
11	K02848	heptose (I) phosphotransferase [EC:2.7.1.-]	Lipopolysaccharide biosynthesis|
12	K11891	type VI secretion system protein ImpL	Bacterial secretion system|
12	K02461	general secretion pathway protein L	Bacterial secretion system|
12	K02452	general secretion pathway protein C	Bacterial secretion system|
12	K02462	general secretion pathway protein M	Bacterial secretion system|
13	K07709	two-component system, NtrC family, sensor histidine kinase HydH [EC:2.7.13.3]	Two-component system|
13	K07711	two-component system, NtrC family, sensor histidine kinase GlrK [EC:2.7.13.3]	Two-component system|
13	K07677	two-component system, NarL family, capsular synthesis sensor histidine kinase RcsC [EC:2.7.13.3]	Two-component system|
13	K07673	two-component system, NarL family, nitrate/nitrite sensor histidine kinase NarX [EC:2.7.13.3]	Two-component system|
13	K07787	Cu(I)/Ag(I) efflux system membrane protein CusA/SilA	Two-component system|
13	K07687	two-component system, NarL family, captular synthesis response regulator RcsB	Two-component system|
13	K07784	MFS transporter, OPA family, hexose phosphate transport protein UhpT	Two-component system|
13	K07640	two-component system, OmpR family, sensor histidine kinase CpxA [EC:2.7.13.3]	Two-component system|
13	K07675	two-component system, NarL family, sensor histidine kinase UhpB [EC:2.7.13.3]	Two-component system|
13	K05874	methyl-accepting chemotaxis protein I, serine sensor receptor	Two-component system|Bacterial chemotaxis|
13	K03414	chemotaxis protein CheZ	Bacterial chemotaxis|
13	K08348	formate dehydrogenase-N, alpha subunit [EC:1.2.1.2]	Glyoxylate and dicarboxylate metabolism|Methane metabolism|Two-component system|
13	K07647	two-component system, OmpR family, sensor histidine kinase TorS [EC:2.7.13.3]	Two-component system|
13	K07796	Cu(I)/Ag(I) efflux system outer membrane protein CusC/SilC	Two-component system|
13	K03532	trimethylamine-N-oxide reductase (cytochrome c) 1, cytochrome c-type subunit TorC	Two-component system|
13	K07648	two-component system, OmpR family, aerobic respiration control sensor histidine kinase ArcB [EC:2.7.13.3]	Two-component system|
13	K07811	trimethylamine-N-oxide reductase (cytochrome c) 1 [EC:1.7.2.3]	Two-component system|
13	K00371	nitrate reductase beta subunit [EC:1.7.99.4]	Nitrogen metabolism|Two-component system|
13	K04013	cytochrome c-type protein NrfB	Nitrogen metabolism|
13	K07701	two-component system, CitB family, sensor histidine kinase DcuS [EC:2.7.13.3]	Two-component system|
13	K07666	two-component system, OmpR family, response regulator QseB	Two-component system|
13	K07797	multidrug resistance protein K	Two-component system|
13	K07789	RND superfamily, multidrug transport protein MdtC	Two-component system|
13	K06080	RcsF protein	Two-component system|
13	K10001	glutamate/aspartate transport system substrate-binding protein	ABC transporters|Two-component system|
13	K10110	maltose/maltodextrin transport system permease protein	ABC transporters|
13	K06073	vitamin B12 transport system permease protein	ABC transporters|
13	K10831	taurine transport system ATP-binding protein [EC:3.6.3.36]	ABC transporters|
13	K10111	maltose/maltodextrin transport system ATP-binding protein [EC:3.6.3.19]	ABC transporters|

13	K07643	two-component system, OmpR family, sensor histidine kinase BasS [EC:2.7.13.3]	Two-component system|
13	K02403	flagellar transcriptional activator FlhD	Two-component system|Flagellar assembly|
13	K07676	two-component system, NarL family, sensor histidine kinase RcsD [EC:2.7.13.3]	Two-component system|
13	K07688	two-component system, NarL family, response regulator, fimbrial Z protein, FimZ	Two-component system|
13	K06858	vitamin B12 transport system substrate-binding protein	ABC transporters|
13	K09475	outer membrane pore protein C	Two-component system|
13	K10000	arginine transport system ATP-binding protein [EC:3.6.3.-]	ABC transporters|
13	K06074	vitamin B12 transport system ATP-binding protein [EC:3.6.3.33]	ABC transporters|
13	K10014	histidine transport system substrate-binding protein	ABC transporters|
13	K06159	putative ATP-binding cassette transporter	ABC transporters|
13	K10017	histidine transport system ATP-binding protein [EC:3.6.3.21]	ABC transporters|
13	K02567	periplasmic nitrate reductase NapA [EC:1.7.99.4]	Nitrogen metabolism|
13	K02569	cytochrome c-type protein NapC	Nitrogen metabolism|
13	K03314	Na+:H+ antiporter, NhaB family	Methane metabolism|
13	K10555	AI-2 transport system substrate-binding protein	ABC transporters|
13	K02394	flagellar P-ring protein precursor FlgI	Flagellar assembly|
13	K10004	glutamate/aspartate transport system ATP-binding protein [EC:3.6.3.-]	ABC transporters|Two-component system|
13	K07674	two-component system, NarL family, nitrate/nitrite sensor histidine kinase NarQ [EC:2.7.13.3]	Two-component system|
13	K00373	nitrate reductase delta subunit	Nitrogen metabolism|Two-component system|
13	K01637	isocitrate lyase [EC:4.1.3.1]	Glyoxylate and dicarboxylate metabolism|
13	K07664	two-component system, OmpR family, response regulator BaeR	Two-component system|
13	K07702	two-component system, CitB family, response regulator CitB	Two-component system|
13	K02391	flagellar basal-body rod protein FlgF	Flagellar assembly|
13	K00374	nitrate reductase gamma subunit [EC:1.7.99.4]	Nitrogen metabolism|Two-component system|
13	K09999	arginine transport system permease protein	ABC transporters|
13	K01682	aconitate hydratase 2 / 2-methylisocitrate dehydratase [EC:4.2.1.3 4.2.1.99]	Citrate cycle (TCA cycle)|Glyoxylate and dicarboxylate metabolism|Propanoate metabolism|Carbon fixation pathways in prokaryotes|
13	K00163	pyruvate dehydrogenase E1 component [EC:1.2.4.1]	Glycolysis / Gluconeogenesis|Citrate cycle (TCA cycle)|Pyruvate metabolism|Butanoate metabolism|
13	K04021	aldehyde dehydrogenase	Benzoate degradation|Pyruvate metabolism|Dioxin degradation|Xylene degradation|
13	K00529	ferredoxin--NAD+ reductase [EC:1.18.1.3]	Fatty acid metabolism|Phenylalanine metabolism|Chlorocyclohexane and chlorobenzene degradation|Benzoate degradation|Dioxin degra... <Preview truncated at 128 characters>
13	K01093	4-phytase / acid phosphatase [EC:3.1.3.26 3.1.3.2]	Inositol phosphate metabolism|Aminobenzoate degradation|Riboflavin metabolism|
13	K03782	catalase-peroxidase [EC:1.11.1.21]	Phenylalanine metabolism|Tryptophan metabolism|Phenylpropanoid biosynthesis|
13	K02611	ring-1,2-phenylacetyl-CoA epoxidase subunit PaaC [EC:1.14.13.149]	Phenylalanine metabolism|
13	K05708	3-phenylpropionate/cinnamic acid dioxygenase subunit alpha [EC:1.14.12.19]	Phenylalanine metabolism|
13	K01782	3-hydroxyacyl-CoA dehydrogenase / enoyl-CoA hydratase / 3-hydroxybutyryl-CoA epimerase [EC:1.1.1.35 4.2.1.17 5.1.2.3]	Fatty acid metabolism|Valine, leucine and isoleucine degradation|Geraniol degradation|Lysine degradation|Tryptophan metabolism|b... <Preview truncated at 128 characters>
13	K03841	fructose-1,6-bisphosphatase I [EC:3.1.3.11]	Glycolysis / Gluconeogenesis|Pentose phosphate pathway|Fructose and mannose metabolism|Methane metabolism|Carbon fixation in pho... <Preview truncated at 128 characters>
13	K00117	quinoprotein glucose dehydrogenase [EC:1.1.5.2]	Pentose phosphate pathway|
13	K01690	phosphogluconate dehydratase [EC:4.2.1.12]	Pentose phosphate pathway|
13	K00855	phosphoribulokinase [EC:2.7.1.19]	Carbon fixation in photosynthetic organisms|

13	K12660	2-dehydro-3-deoxy-L-rhamnonate aldolase [EC:4.1.2.-]	Fructose and mannose metabolism|
13	K12661	L-rhamnonate dehydratase [EC:4.2.1.90]	Fructose and mannose metabolism|
13	K16013	ATP-binding cassette, subfamily C, bacterial CydD	ABC transporters|
13	K04022	alcohol dehydrogenase	Glycolysis / Gluconeogenesis|
13	K02753	PTS system, arbutin-, cellobiose-, and salicin-specific IIC component	Glycolysis / Gluconeogenesis|Phosphotransferase system (PTS)|
13	K10985	PTS system, galactosamine-specific IIC component	Galactose metabolism|Phosphotransferase system (PTS)|
13	K12112	evolved beta-galactosidase subunit beta	Galactose metabolism|Other glycan degradation|
13	K12111	evolved beta-galactosidase subunit alpha [EC:3.2.1.23]	Galactose metabolism|Other glycan degradation|
13	K08484	phosphotransferase system, enzyme I, PtsP [EC:2.7.3.9]	Phosphotransferase system (PTS)|
13	K02774	PTS system, galactitol-specific IIB component [EC:2.7.1.69]	Galactose metabolism|Phosphotransferase system (PTS)|
13	K02752	PTS system, arbutin-, cellobiose-, and salicin-specific IIB component [EC:2.7.1.69]	Glycolysis / Gluconeogenesis|Phosphotransferase system (PTS)|
13	K01061	carboxymethylenebutenolidase [EC:3.1.1.45]	Chlorocyclohexane and chlorobenzene degradation|Fluorobenzoate degradation|Toluene degradation|
13	K01070	S-formylglutathione hydrolase [EC:3.1.2.12]	Methane metabolism|
13	K05711	2,3-dihydroxy-2,3-dihydrophenylpropionate dehydrogenase [EC:1.3.1.87]	Phenylalanine metabolism|
13	K00932	propionate kinase [EC:2.7.2.15]	Propanoate metabolism|
13	K03776	aerotaxis receptor	Two-component system|Bacterial chemotaxis|
13	K01659	2-methylcitrate synthase [EC:2.3.3.5]	Propanoate metabolism|
13	K07788	RND superfamily, multidrug transport protein MdtB	Two-component system|
13	K02618	oxepin-CoA hydrolase / 3-oxo-5,6-dehydrosuberyl-CoA semialdehyde dehydrogenase [EC:3.3.2.12 1.17.1.7]	Phenylalanine metabolism|
13	K03777	D-lactate dehydrogenase [EC:1.1.1.28]	Pyruvate metabolism|
13	K06445	acyl-CoA dehydrogenase [EC:1.3.99.-]	Fatty acid metabolism|
13	K10016	histidine transport system permease protein	ABC transporters|
13	K05875	methyl-accepting chemotaxis protein II, aspartate sensor receptor	Two-component system|Bacterial chemotaxis|
14	K00322	NAD(P) transhydrogenase [EC:1.6.1.1]	Nicotinate and nicotinamide metabolism|
14	K11751	5-nucleotidase / UDP-sugar diphosphatase [EC:3.1.3.5 3.6.1.45]	Purine metabolism|Pyrimidine metabolism|Nicotinate and nicotinamide metabolism|
14	K05851	adenylate cyclase, class 1 [EC:4.6.1.1]	Purine metabolism|Vibrio cholerae pathogenic cycle|
14	K01139	guanosine-3,5-bis(diphosphate) 3-pyrophosphohydrolase [EC:3.1.7.2]	Purine metabolism|
14	K09018	pyrimidine oxygenase [EC:1.14.99.46]	Pyrimidine metabolism|
14	K00892	inosine kinase [EC:2.7.1.73]	Purine metabolism|
14	K10939	accessory colonization factor AcfD	Vibrio cholerae pathogenic cycle|
14	K00073	ureidoglycolate dehydrogenase [EC:1.1.1.154]	Purine metabolism|
15	K09470	gamma-glutamylputrescine synthase [EC:6.3.1.11]	Arginine and proline metabolism|
15	K00840	succinylornithine aminotransferase [EC:2.6.1.81]	Arginine and proline metabolism|
15	K09472	gamma-glutamyl-gamma-aminobutyraldehyde dehydrogenase [EC:1.2.1.-]	Arginine and proline metabolism|
15	K06447	succinylglutamic semialdehyde dehydrogenase [EC:1.2.1.71]	Arginine and proline metabolism|
15	K05526	succinylglutamate desuccinylase [EC:3.5.1.96]	Arginine and proline metabolism|
15	K09471	gamma-glutamylputrescine oxidase [EC:1.4.3.-]	Arginine and proline metabolism|
16	K00124	formate dehydrogenase, beta subunit	Glyoxylate and dicarboxylate metabolism|Methane metabolism|
16	K08349	formate dehydrogenase-N, beta subunit	Glyoxylate and dicarboxylate metabolism|Methane metabolism|Two-component system|
16	K07639	two-component system, OmpR family, sensor histidine kinase RstB [EC:2.7.13.3]	Two-component system|
16	K07773	two-component system, OmpR family, aerobic respiration control protein ArcA	Two-component system|
16	K07686	two-component system, NarL family, uhpT operon response regulator UhpA	Two-component system|
16	K07708	two-component system, NtrC family, nitrogen regulation sensor histidine kinase GlnL [EC:2.7.13.3]	Two-component system|
16	K07670	two-component system, OmpR family, response regulator MtrA	Two-component system|
16	K09476	outer membrane pore protein F	Two-component system|

16	K10914	CRP/FNR family transcriptional regulator, cyclic AMP receptor protein	Two-component system|Vibrio cholerae pathogenic cycle|
16	K11472	glycolate oxidase FAD binding subunit	Glyoxylate and dicarboxylate metabolism|
17	K10013	lysine/arginine/ornithine transport system substrate-binding protein	ABC transporters|
17	K11076	putrescine transport system ATP-binding protein	ABC transporters|
17	K13892	glutathione transport system ATP-binding protein	ABC transporters|
17	K09998	arginine transport system permease protein	ABC transporters|
17	K02064	thiamine transport system substrate-binding protein	ABC transporters|
18	K10023	arginine/ornithine transport system permease protein	ABC transporters|
18	K02196	heme exporter protein D	ABC transporters|
18	K09997	arginine transport system substrate-binding protein	ABC transporters|
18	K10108	maltose/maltodextrin transport system substrate-binding protein	ABC transporters|Bacterial chemotaxis|
19	K07781	LuxR family transcriptional regulator, capsular biosynthesis positive transcription factor	Two-component system|
19	K07715	two-component system, NtrC family, response regulator GlrR	Two-component system|
19	K10941	sigma-54 specific transcriptional regulator, flagellar regulatory protein A	Two-component system|Vibrio cholerae pathogenic cycle|
19	K07703	two-component system, CitB family, response regulator DcuR	Two-component system|
19	K00404	cytochrome c oxidase cbb3-type subunit I [EC:1.9.3.1]	Oxidative phosphorylation|Two-component system|
19	K10942	two-component system, sensor histidine kinase FlrB [EC:2.7.13.3]	Two-component system|Vibrio cholerae pathogenic cycle|
19	K08478	phosphoglycerate transport regulatory protein PgtC	Two-component system|
19	K08476	two-component system, NtrC family, phosphoglycerate transport system response regulator PgtA	Two-component system|
19	K03533	TorA specific chaperone	Two-component system|
19	K07710	two-component system, NtrC family, sensor histidine kinase AtoS [EC:2.7.13.3]	Two-component system|
19	K08475	two-component system, NtrC family, phosphoglycerate transport system sensor histidine kinase PgtB [EC:2.7.13.3]	Two-component system|
19	K07679	two-component system, NarL family, sensor histidine kinase EvgS [EC:2.7.13.3]	Two-component system|Pertussis|
19	K09477	citrate:succinate antiporter	Two-component system|
19	K07700	two-component system, CitB family, cit operon sensor histidine kinase CitA [EC:2.7.13.3]	Two-component system|
19	K07689	two-component system, NarL family, invasion response regulator UvrY	Two-component system|
19	K07685	two-component system, NarL family, nitrate/nitrite response regulator NarP	Two-component system|
19	K12973	palmitoyl transferase [EC:2.3.1.-]	Pertussis|
20	K14121	energy-converting hydrogenase B subunit L	Methane metabolism|
20	K00440	coenzyme F420 hydrogenase alpha subunit [EC:1.12.98.1]	Methane metabolism|
20	K14101	energy-converting hydrogenase A subunit J	Methane metabolism|
20	K14095	energy-converting hydrogenase A subunit D	Methane metabolism|
20	K11260	formylmethanofuran dehydrogenase subunit G [EC:1.2.99.5]	Methane metabolism|
20	K14113	energy-converting hydrogenase B subunit D	Methane metabolism|
21	K07786	MFS transporter, multidrug resistance protein Y	Two-component system|
21	K10002	glutamate/aspartate transport system permease protein	ABC transporters|Two-component system|
21	K08350	formate dehydrogenase-N, gamma subunit	Glyoxylate and dicarboxylate metabolism|Methane metabolism|Two-component system|
21	K11631	bacitracin transport system ATP-binding protein	ABC transporters|Two-component system|Staphylococcus aureus infection|
21	K07782	LuxR family transcriptional regulator	Two-component system|
21	K17205	putative xylitol transport system substrate-binding protein	ABC transporters|
21	K12972	gyoxylate/hydroxypyruvate reductase A [EC:1.1.1.79 1.1.1.81]	Glycine, serine and threonine metabolism|Pyruvate metabolism|Glyoxylate and dicarboxylate metabolism|
21	K09996	arginine transport system substrate-binding protein	ABC transporters|

22	K04345	protein kinase A [EC:2.7.11.11]	MAPK signaling pathway|Calcium signaling pathway|Chemokine signaling pathway|Meiosis - yeast|Oocyte meiosis|Apoptosis|Vascular s... <Preview truncated at 128 characters>
22	K09047	cyclic AMP-responsive element-binding protein 5	PI3K-Akt signaling pathway|TNF signaling pathway|Cholinergic synapse|Dopaminergic synapse|Insulin secretion|Estrogen signaling p... <Preview truncated at 128 characters>
22	K09487	heat shock protein 90kDa beta	Protein processing in endoplasmic reticulum|PI3K-Akt signaling pathway|NOD-like receptor signaling pathway|Plant-pathogen intera... <Preview truncated at 128 characters>
22	K14012	UBX domain-containing protein 1	Protein processing in endoplasmic reticulum|
23	K03788	acid phosphatase (class B) [EC:3.1.3.2]	Aminobenzoate degradation|Riboflavin metabolism|
23	K01913	NONE	Aminobenzoate degradation|Propanoate metabolism|Limonene and pinene degradation|Caprolactam degradation|Tropane, piperidine and ... <Preview truncated at 128 characters>
23	K01825	3-hydroxyacyl-CoA dehydrogenase / enoyl-CoA hydratase / 3-hydroxybutyryl-CoA epimerase / enoyl-CoA isomerase [EC:1.1.1.35 4.2.1.... <Preview truncated at 128 characters>	Fatty acid metabolism|Valine, leucine and isoleucine degradation|Geraniol degradation|Lysine degradation|Tryptophan metabolism|b... <Preview truncated at 128 characters>
23	K10806	acyl-CoA thioesterase YciA [EC:3.1.2.-]	Biosynthesis of unsaturated fatty acids|
23	K00276	primary-amine oxidase [EC:1.4.3.21]	Glycine, serine and threonine metabolism|Tyrosine metabolism|Phenylalanine metabolism|beta-Alanine metabolism|Isoquinoline alkal... <Preview truncated at 128 characters>
23	K02612	ring-1,2-phenylacetyl-CoA epoxidase subunit PaaD	Phenylalanine metabolism|
23	K00830	alanine-glyoxylate transaminase / serine-glyoxylate transaminase / serine-pyruvate transaminase [EC:2.6.1.44 2.6.1.45 2.6.1.51]	Alanine, aspartate and glutamate metabolism|Glycine, serine and threonine metabolism|Glyoxylate and dicarboxylate metabolism|Met... <Preview truncated at 128 characters>
23	K03821	polyhydroxyalkanoate synthase [EC:2.3.1.-]	Butanoate metabolism|
23	K00137	aminobutyraldehyde dehydrogenase [EC:1.2.1.19]	Arginine and proline metabolism|beta-Alanine metabolism|
23	K05713	2,3-dihydroxyphenylpropionate 1,2-dioxygenase [EC:1.13.11.16]	Phenylalanine metabolism|
24	K00580	tetrahydromethanopterin S-methyltransferase subunit D [EC:2.1.1.86]	Methane metabolism|
24	K00205	formylmethanofuran dehydrogenase subunit F [EC:1.2.99.5]	Methane metabolism|
24	K14094	energy-converting hydrogenase A subunit C	Methane metabolism|
24	K03421	methyl-coenzyme M reductase subunit C	Methane metabolism|
25	K00232	acyl-CoA oxidase [EC:1.3.3.6]	Fatty acid metabolism|alpha-Linolenic acid metabolism|Biosynthesis of unsaturated fatty acids|PPAR signaling pathway|Peroxisome|
25	K00273	D-amino-acid oxidase [EC:1.4.3.3]	Glycine, serine and threonine metabolism|Penicillin and cephalosporin biosynthesis|Arginine and proline metabolism|D-Arginine an... <Preview truncated at 128 characters>
25	K17217	cystathionine gamma-lyase / homocysteine desulfhydrase [EC:4.4.1.1 4.4.1.2]	Glycine, serine and threonine metabolism|Cysteine and methionine metabolism|Sulfur metabolism|
25	K00619	amino-acid N-acetyltransferase [EC:2.3.1.1]	Arginine and proline metabolism|
26	K11004	ATP-binding cassette, subfamily B, bacterial HlyB/CyaB	ABC transporters|Bacterial secretion system|Pertussis|
26	K16202	dipeptide transport system ATP-binding protein	ABC transporters|
26	K09696	sodium transport system permease protein	ABC transporters|Two-component system|
26	K02660	twitching motility protein PilJ	Two-component system|
27	K07653	two-component system, OmpR family, sensor histidine kinase MprB [EC:2.7.13.3]	Two-component system|
27	K07683	two-component system, NarL family, sensor histidine kinase NreB [EC:2.7.13.3]	Two-component system|
27	K02658	twitching motility two-component system response regulator PilH	Two-component system|
27	K11382	MFS transporter, OPA family, phosphoglycerate transporter protein	Two-component system|

27	K00411	ubiquinol-cytochrome c reductase iron-sulfur subunit [EC:1.10.2.2]	Oxidative phosphorylation|Nitrogen metabolism|Two-component system|Cardiac muscle contraction|Alzheimers disease|Parkinsons di... <Preview truncated at 128 characters>
27	K02261	cytochrome c oxidase subunit 2	Oxidative phosphorylation|Cardiac muscle contraction|Alzheimers disease|Parkinsons disease|Huntingtons disease|
27	K08479	two-component system, OmpR family, clock-associated histidine kinase SasA [EC:2.7.13.3]	Two-component system|
27	K03881	NADH-ubiquinone oxidoreductase chain 4 [EC:1.6.5.3]	Oxidative phosphorylation|Parkinsons disease|
27	K01674	carbonic anhydrase [EC:4.2.1.1]	Nitrogen metabolism|
27	K03889	ubiquinol-cytochrome c reductase cytochrome c subunit	Oxidative phosphorylation|
28	K02204	homoserine kinase type II [EC:2.7.1.39]	Glycine, serine and threonine metabolism|
28	K01758	cystathionine gamma-lyase [EC:4.4.1.1]	Glycine, serine and threonine metabolism|Cysteine and methionine metabolism|Selenocompound metabolism|
28	K08964	methylthioribulose-1-phosphate dehydratase [EC:4.2.1.109]	Cysteine and methionine metabolism|
28	K10764	O-succinylhomoserine sulfhydrylase [EC:2.5.1.-]	Cysteine and methionine metabolism|
29	K09693	teichoic acid transport system ATP-binding protein [EC:3.6.3.40]	ABC transporters|
29	K11708	manganese/zinc/iron transport system permease protein	ABC transporters|
29	K11632	bacitracin transport system permease protein	ABC transporters|Two-component system|Staphylococcus aureus infection|
29	K11629	two-component system, OmpR family, bacitracin resistance sensor histidine kinase BceS [EC:2.7.13.3]	Two-component system|
29	K06375	stage 0 sporulation protein B (sporulation initiation phosphotransferase) [EC:2.7.-.-]	Two-component system|
29	K10554	fructose transport system ATP-binding protein	ABC transporters|

* Modules with universal presence (modules 1-3) were omitted from this table due to space constraints.

Appendix C – Supplementary material for chapter 4

C.1 Supplementary Methods

C.1.1 Evaluating reference genome cluster definitions and read mappability

The 260 reference genomes were assigned to 101 clusters, according to sequence similarity of 40 marker genes [124]. The clustering was performed in a previous study, with clusters serving as a proxy for species and individual genomes within a cluster representing instances of intra-species genomic variation. Clusters ranged in size; many clusters contained just one genome, while the largest cluster contained 63 genomes. Clusters could contain genomes from a single taxonomic clade or several clades, though most clusters agreed with current species definitions.

To validate that these clusters were suitable for our mapping pipeline, we performed multiple simulation-based analyses. Specifically, we aimed to examine whether short reads that originate from a given genome and a given gene map to the correct genome cluster and to the correct KO. Notably, such reads are not necessarily required to map solely to the genome from which they originated (as this genome will often not be available in the reference genomes database) nor to the exact gene they originated from. Rather, reads should map to *some* genome (or genomes) from the same cluster, and to gene regions with the same KO annotation. Moreover, for our pipeline to correctly estimate gene copy numbers, mapping should also be robust to sequencing errors and should correctly exclude reads originating from genomes not represented by any of the clusters included in our analysis.

To this end, custom perl scripts were used to simulate reads by extracting randomly selected stretches of 75 base pairs from the KO-annotated gene regions of 10 query genomes from 8 different genome clusters (*Bifidobacterium longum NCC2705, Streptococcus mitis B6, Bacteroides ovatus ATCC 8483, Bacteroides vulgatus ATCC 8482, Escherichia coli SMS-3-5, Alistipes putredinis DSM 17216, Citrobacter youngae ATCC 29220, Eubacterium rectale ATCC 33656, Prevotella copri DSM 18205, Bacteroides vulgatus PC510*). Simulated reads were then aligned concurrently to the set of reference genomes, with a maximum allowable edit distance of 5 and up to 75 tied best alignments reported (see also next section, "*Validating maximum edit distance for read alignment"*). Alignment results were parsed to bin each read according to the cluster and KO from which the read originated (query KC) and the cluster and KO to which it mapped best. Specifically, reads could be unmapped, mapped to >75 different regions, mapped to a genome from the correct cluster or mapped to a genome from an incorrect cluster, and reads could be mapped to a region associated with the correct KO, an incorrect KO, an unannotated gene region, or an intergenic region. Reads mapping to multiple regions were given fractional counts distributed evenly across the set of corresponding KCs. For regions with

multiple annotations, if any of the query KOs matched any of the target KOs, the target was considered to be the ‘right KO’.

We first mapped 45,855 simulated KO-annotated reads from the 10 query genomes above to the full set of 260 reference genomes (which includes the 10 query genomes). As expected, each read correctly mapped to the genomic region from which it originated. Clearly, however, many reads mapped equally well to other regions. When distributing read counts across all tied alignments as described above, we found that 62.1% of fractional counts were assigned to the correct KO in the original genome, 36.2% were assigned to the correct KO in a different genome from the same cluster, and only 1.7% of fractional counts were incorrectly assigned (Fig. C.1A); 0.4% were assigned to the wrong cluster, while 1.3% were not assigned to any cluster (either because the aligned region was intergenic or unannotated, or the read mapped to>75 regions). This finding suggests that the cluster definitions and parameters used allow reads to map uniquely to the genome of origin or to an identical region from another genome in the same genome cluster, and that such identical regions are only rarely found in another genome cluster.

To assess the effect of short-read sequencing errors, we next applied a position-dependent error profile created with Ibis [165] from an Illumina sequencing run, uniformly magnified with custom perl scripts to achieve 1.5% error rate. These error-adjusted reads were then aligned to the set of 260 reference genomes, as above. Evidently, the addition of an error model did not markedly change the mapping accuracy observed above (and none of the read mapping statistics reported above changed by more than 1%; Fig. C.1B). Again, relatively few reads remained unmapped (e.g., reads assigned to the ‘no cluster’ bin, which may now also include reads that were unmapped due to sequencing errors), and reads from each genome were still far more likely to be aligned to regions within the correct genome cluster rather than regions in another genome cluster.

As noted above, a primary assumption of our read alignment pipeline is that reads from a strain which would group with one of the clusters in our database but for which a reference genome is not yet available, will still map to a reference genome within the correct cluster. This allows us to detect novel intra-cluster variation at regions of altered coverage. We further assume that such reads will map to regions from the same orthologous group of genes, as defined by KEGG (KOs). To validate these assumptions, we re-aligned the error-adjusted reads from the above simulation to the reference database, but now, when aligning each read, we removed the genome of origin from the database. Overall, we found that among reads for which the query KC was present in the database, 66.8% of fractional counts were correctly mapped to the same KC as the query, while 20.8% were incorrectly unmapped, and only 0.8% were mapped to the wrong KC (Fig. C.1C). In some cases however, removing the genome of origin resulted in a reference database in which the correct KC was no longer present – either because the removed genome was the only one in the cluster, or because no other genome in the cluster contained the KO. In these cases, we defined an unmapped read as 'correctly unmapped', while a read mapping to any other KC was defined as 'incorrectly mapped'. Among reads for which the query KC was no longer present, 98.9% were correctly unmapped, and 1.1% were mapped to another KC. These findings indicate that the specificity of our pipeline is high; even when the genome of origin was removed from the database, reads mostly aligned to the query KC when it was present, and were almost always unmapped when it was not. Notably, a significant number of reads remained unmapped at a maximum edit distance of 5 when a correct KC was present. However, most of these reads came from 2 specific genomes, while the false negative rate in the other 8 genomes was very low. As noted below, we address extreme cases of genomes with consistently false mapping by filtering the set of reference genomes and removing genomes with high mapping error rates.

To determine whether these trends hold true on a more global scale, below we additionally examined simulated reads from all 260 reference genomes (see 'Determining mapping error rates and filtering clusters and KOs').

C.1.2 Validating maximum edit distance for read alignment

Since edit distance was used as the primary threshold for short read alignments, we additionally performed a simulation-based analysis to confirm that a maximum edit distance (MED) of 5 would allow reads to be aligned uniquely to the correct KC, while minimizing both the number of unmapped reads and incorrectly mapped reads. For this simulation, we again mapped the error-adjusted reads simulated from all 260 genomes to a reference database in which the genome of origin had been removed as described above, but this time allowed best alignments at a range of MEDs from 0 to 10. We then examined changes in mapping accuracy over this range of MEDs (Fig. C.1D). We found that at all MEDs greater than 0, the majority of reads were either correctly mapped or correctly unmapped (ie., when the query KC was no longer in the reference database). The number of correctly mapped reads increased rapidly from a MED of 0 to a MED of 5, and remained relatively stable at higher MEDs. Notably, the number of incorrectly mapped reads continued to increase over the entire range of MEDs tested, suggesting that a MED much higher than 5 should not be used. Among reads for which the KC was present, the major source of erroneous mappings was to unannotated regions in the correct cluster (Fig. C.1D-inset). This may imply that the correct KO in fact exists in this cluster, but has not been correctly annotated as a gene, or has perhaps lost its functionality and become a pseudogene. Though these mapping errors stabilized at MEDs>5, the rate of incorrect mappings to the right KO in the wrong cluster continued to increase both for reads for which the KC was present in the database, as well as those for which it was absent. In light of

the above analysis, a MED of 5 was used in the alignment of all sample data to the 260 reference genomes.

C.1.3 Determining mapping error rates and filtering reference genomes and KOs

To confirm that the mapping accuracy observed above for the 10 query genomes and the mapping parameters optimized in the previous section for the read alignment pipeline apply on a more global scale, we simulated reads from *all* 260 reference genomes and repeated the analysis described above. We found that the majority of reads still mapped to the correct KC when it was present in the database (65.1%), and correctly remained unmapped when it was not (23.1%), with a total error rate (incorrectly mapped + incorrectly unmapped) of only 11.8%.

To further improve the accuracy of our pipeline, we additionally examined whether there were a small number of genomes or KOs which were especially prone to incorrect mapping and that contributed disproportionally to observed inaccuracies, potentially due to various evolutionary and technical factors. We therefore assessed the accuracy of our pipeline for each of the 260 genomes (Fig. C.2A) and each of the 4,304 KOs from which at least 100 reads had been simulated (Fig. C.2C). Specifically, we used the simulations described above and calculated a mapping error rate for each genome and each KO, defined as the percent of simulated reads originating from the KO or genome that were incorrectly unmapped or incorrectly mapped. We identified 25 genomes with error rates >40% (Fig. C.2B). Excluding these genomes from our analysis, we find that the overall genome-wide error rate is reduced by nearly half; among the remaining set of 235 genomes, 68.7% of reads were correctly mapped, while 25.1% of reads were correctly unmapped, and only 6.2% of reads were incorrect (1.4% incorrectly mapped; 4.8% incorrectly unmapped). The rest of the analysis was carried out with this filtered set of 235 genomes, corresponding to 96 clusters. Error rates for the KOs varied greatly, however the vast majority (4,272 KOs) had an error rate <=50%. Most of these errors

were due to incorrectly *unmapped* reads and only 8.5% of KOs had any incorrectly *mapped* reads. These errors could conceivably be due to either misannotation or low intra-species sequence conservation, among other factors. For the rest of our analysis, we focused only on KOs with a combined error rate <=50%, excluding the 35 KOs with a higher error rate (Fig. C.2D).

C.1.4 Identifying and validating marker KOs

We set out to identify a set of marker KOs whose coverage could be used as a proxy for the abundance of each genome cluster in each sample. Ideally, each of these KOs would be present in exactly one copy in each reference genome (high universality), would have a low mapping error rate in our simulated alignments (high alignment accuracy), and would have consistent relative coverage by reads from any given metagenomic sample (high coverage consistency). We accordingly obtained the set of 40 marker COGs used by Schloissnig *et al.* [124], translated COG annotations to KO annotations using a KEGG-generated mapping file (http://www.genome.jp/files/ko2cog.xl), and filtered the associated KOs to a smaller set based on the three criteria described above. Specifically, we defined universality as the percent of reference genomes (out of 260) in which the KO had a copy number >=1. We defined alignment accuracy as 1 minus the KO mapping error rate (see *Determining mapping error rates and filtering genomes and KOs,* above). To assess coverage consistency, we first summed the coverage of each KO in each sample across all clusters, normalized by the mean within each sample, and recorded the distance between these values and 1. For each KO, coverage consistency was defined as 1 minus the average across all samples. We filtered the 40 KOs to identify those with universality>0.95, alignment accuracy >0.9, and coverage consistency >0.85 (Fig. C.3A). 13 KOs met all three criteria (Fig. C.3B) and were used in the final analysis as marker KOs for calculation of cluster abundance (see Experimental Procedures).

C.1.5 Comparison of highly variable and set-specific variable KCs to known strain variation

To verify our data processing pipeline, we examined the overlap between KCs identified as variable across samples by our analysis and KCs that vary in copy number across reference genomes in our database. As described in the main text, overall, this overlap was very high (80.9% for highly variable KCs, 70.4% for set-specific variable KCs). To ensure however that this high overlap was not due to some detection bias stemming from the use of these reference genomes in our pipeline, we wished to confirm that a similar overlap can also be observed when comparing our predicted variation to variation found between genomes not included in our database.

We therefore first identified three single-genome clusters - cluster 22 (*Dorea longicatena*), cluster 23 (*Ruminococcus lactaris*), and cluster 34 (*Dorea formicigenerans*) – each of which could be associated with an additional annotated reference genome from IMG [55] representing a different strain from this cluster (*Dorea longicatena AGR2136*, *Dorea formicigenerans 4_6_53AFAA*, *Ruminococcus lactaris CC59_002D*). These ‘new’ genomes were not included in our reference database and were therefore not used as targets in the read alignment process. For consistency, we downloaded from IMG the KO annotations for both the ‘new’ genomes and for the three corresponding ‘reference’ genomes already in our database, and limited our analysis to KCs for which IMG annotations for the reference genomes agreed with the annotations in our database. We also examined 44 newly-sequenced reference strains from the NCBI database that were sequenced after our initial analysis, and were therefore not included in the original alignment and annotation pipeline. We annotated each of these additional genomes using the same KEGG BLAST pipeline as with the main reference set.

We compared the KCs identified as variable by our analysis with KCs that vary in copy number between the reference genomes used for mapping and the newly obtained genomes

from either IMG or NCBI. Examining the variation present in the IMG genomes, we find high overlap with detected variable KCs, with 71%, 64%, and 39% of the KCs that were identified as highly variable across samples by our analysis in clusters 22, 34 and 23 respectively corresponding to KCs that vary in copy number between the reference genome and the new genome (Fig. 4.4B). Importantly, these values are comparable to the overlap observed in the 4 genomes clusters in our database in which two reference genomes were included as alignment targets (mean 63%), suggesting that variation detected by our pipeline was not unduly influenced by the specific strains used as references during read alignment. When examining set-specific variable KCs, this overlap was still high (47%, 45%, and 28% for the three clusters respectively), yet as demonstrated for other clusters, identified variable KCs further included many instances of novel variation (Fig. C.5B). Examining the additional genomes from NCBI (Figs. 4.4C and C.5C) we find 302 instances in which copy number variation detected in our samples (including 39 highly variable KCs and 263 set-specific variable KCs) was reflected in copy number differences in these additional sequenced reference genomes. In the two cases in which additional genomes were examined for clusters that originally were represented by only a single reference, over 70% of the detected highly variable KCs, and close to 60% of set-specific variable KCs exhibited copy number differences between the original and additional genome.

C.1.6 Cross-validation of variable KCs

To examine the robustness and sensitivity of our pipeline, we performed a cross-validation analysis, testing whether significantly high variation detected using a subset of our samples is predictive of variation observed in the remaining non-overlapping subset of samples. We focused on the 30 genome clusters that were identified as present in at least 20 samples. For each cluster, the samples containing this cluster were randomly divided into 5 equally-populated cross-validation groups. For each cluster we then performed 5 rounds of highly variable KC detection (as defined by our pipeline), each time leaving out a different sample group (testing

set) and detecting variation only based on the remaining 4 groups combined (training set). We then examined whether KCs detected as highly variable in the training set also exhibited significantly higher variation among samples in the testing set by comparing the median absolute deviation of these genes to the median absolute deviation of KCs not detected as variable and using a t-test to assess significance. We found that across all rounds of cross-validation and in each of the 30 clusters tested, genes detected as highly variable in the training set indeed exhibited higher variation in the testing set, confirming the robustness of our pipeline and demonstrating that high variability observed in the copy number of certain genes is not merely due to some extreme (and potentially spurious) variation in just one or a few samples.

C.1.7 Mock community simulation and analysis

To assess the accuracy of our pipeline and the resolution of our variable KC detection scheme, we created a synthetic dataset of metagenomic samples in which cluster abundances and KC copy numbers were known *a priori*. Specifically, expanding on the simulation procedure described in [104], we generated 40 simulated samples, each of which consisted of 13 million 75-bp reads (comparable to the sample with the lowest sequencing depth in the Danish/Spanish cohort analyzed in our study) extracted at random from a sample-specific community of reference genomes. To generate these samples, 50 reference genomes from 50 different clusters (minimizing confounding variation) were chosen at random to be included in the simulation. For each sample, the community was constructed by randomly assigning a relative abundance (up to 25 fold variation) to each of the 50 reference genomes. We introduced gene-level variation by deleting or duplicating a subset of 50±35 randomly selected genes in each genome, using a probabilistic model that assigned randomly chosen gain and loss rates to each of these genes. 75-bp regions were then extracted from this simulated community of genomes, and subject to a 1.5% sequencing error model (see [104] for more details). We then used our framework to analyze these simulated samples, aligning simulated reads to the original set of

260 reference genomes, calculating species abundances and KC copy numbers as defined by our pipeline, and calling copy number variation.

We compared the obtained species abundances, copy number estimates, and predicted sets of variable KCs to the parameters used to generate the simulated samples in order to quantify the accuracy of our pipeline and its ability to recover species and gene features. As demonstrated in Fig. C.4A, species abundance prediction was extremely accurate with a correlation of 0.993 ($p<10^{-300}$; Pearson correlation test) between predicted and real relative abundance values across the 40 simulated communities and 50 species analyzed, confirming our marker genes-based approach for inferring community composition. Similarly, we confirmed that our copy number estimates correctly recover the copy number of each gene in each genome cluster (Figure C.4B). Copy number estimation accuracy increased with coverage, from 87.6% for genome clusters with low coverage (1x-2x), to 97.8% for clusters with higher coverage (>5x). Estimation accuracy also depended on the underlying copy number, with low copy numbers predicted more accurately than high copy numbers. Overall, the copy number of 96% of KCSs were correctly predicted in 'detectable' genome clusters (coverage >1x as defined by our pipeline). Importantly, overall estimation accuracy dropped to 60.1% for undetectable clusters (coverage <1x), justifying our decision to remove such clusters from downstream analysis. We further examined how many of the KCs in which variation was introduced when simulating the samples were identified as variable by our pipeline. Overall accuracy in detectable clusters was high (98.1%). Sensitivity and specificity were also high (98.8% and 98.1% respectively) though specificity decreased for KCs with high underlying copy numbers (e.g., 81.4% for KCs with copy number 4 and 70.1% for KCs with copy number 5), potentially due to decreased accuracy in copy number estimates reported above and resulting spurious inter-sample variability. Indeed the vast majority of KCs with high median copy number in the dataset analyzed in the main text were detected as variable by our pipeline, and while most of

them likely represent true instances of variation (note, for example, that 77.6% of the KCs with median copy number >=5 vary in copy number among the genomes included in our reference set), our confidence in detecting variability in such KCs may be limited. Importantly, however, such KCs represent a very small fraction of the KCs in this dataset (e.g., only 0.56% of KCs have median copy number >=5). Yet, to confirm that such potentially spurious variable KCs do not affect our findings, we repeated our analysis of variable KCs, filtering out all KCs with median copy number >=5 (see Table C.1). We found that this did not qualitatively change the trends reported in the main text. Specifically, of the functional enrichments reported, 85% (91/107) still held with this filtered set of variable genes (Table C.2). Similarly, the results reported in the main text with regard to individual variable KCs were not affected by this filtering. Finally, we examined whether our pipeline correctly classified KCs as highly vs. set-specific variable, plotting the recall of variable genes and their classification as a function of the percentage of simulated samples in which each KC was deleted or duplicated (Fig. C.4C). This analysis again demonstrated that our pipeline not only successfully recovered the majority of variable KCs but was also able to distinguish between high frequency variation (KCs that vary in many samples) and set-specific variation (KCs that vary in only a small subset of samples).

C.1.8 Analysis of variable functions in an additional sample set from a Chinese cohort

To determine whether trends identified among the functions associated with variable KCs in our original Danish/Spanish cohort extend to other datasets, we applied our copy number variation pipeline to samples from a separate study of the gut microbiomes of Chinese individuals [52]. Mapping parameters and variability detection schemes were identical to those used for the primary dataset. Examining the 73 samples with 75-bp reads from this cohort, we identified 51 genome clusters present in at least one sample. 27 of these clusters were present in at least 10 samples in both datasets, and were assayed for KC variation in the new sample

set, yielding 6,898 highly or set-specific variable KCs. Overall, of the KCs detected as highly variable in the original dataset, 65% (350/538) were identified as highly variable also in the second dataset and 96% (515/538) were identified as either highly or set-specific variable in the second dataset. Of the KCs detected as set-specific variable in the original dataset, 75% (2710/3591) were identified as either highly or set-specific variable in the second dataset. Within each genome cluster, an over-representation analysis was performed to identify the functions associated with the set of variable KCs in each cluster (Table C.2), as described in the analysis of the primary dataset in the main text. Examining the overlap in detected functional classes, 59% (44/74) of the associations reported in the main text for the 27 clusters examined were also found to be significantly associated with copy number variation in the second dataset and this overlap was higher (68%; 17/25) among transport-related functions. Similarly, 67% (10/15) of the functions associated specifically with highly variable genes were also significantly associated with highly variable genes in the second dataset. In certain clusters (ie. *Bacteriodes ovatus* and *Roseburia inulinivorans*), functions found to be over-represented among variable KCs were almost identical in the two datasets.

C.1.9 Mapping rates for metagenomic reads to reference genomes

We mapped a total of 2.47 billion 75bp reads to 260 reference genomes. On average, 34.7% of the reads in each sample could be mapped to a reference genome at an edit distance ≤ 5, although in some samples the mapping rate was as high as 71.5%. This average mapping rate is comparable to the one observed (31%) in mapping these reads to 194 gut-associated genomes in the original study that generated these reads [130], as well as to mapping rates observed in similar studies [52], [124]. These rates are also not surprising given the complexity of gut-associated communities and the predicted prevalence of rare and uncharacterized species. Of the mapped reads, an average of 82.6% overlapped a gene coding region (of which

over a third are annotated with a KO), which is in close agreement with the percentage of the total length of gene coding regions within the genomes in our reference database.

C.2 Supplementary Figures

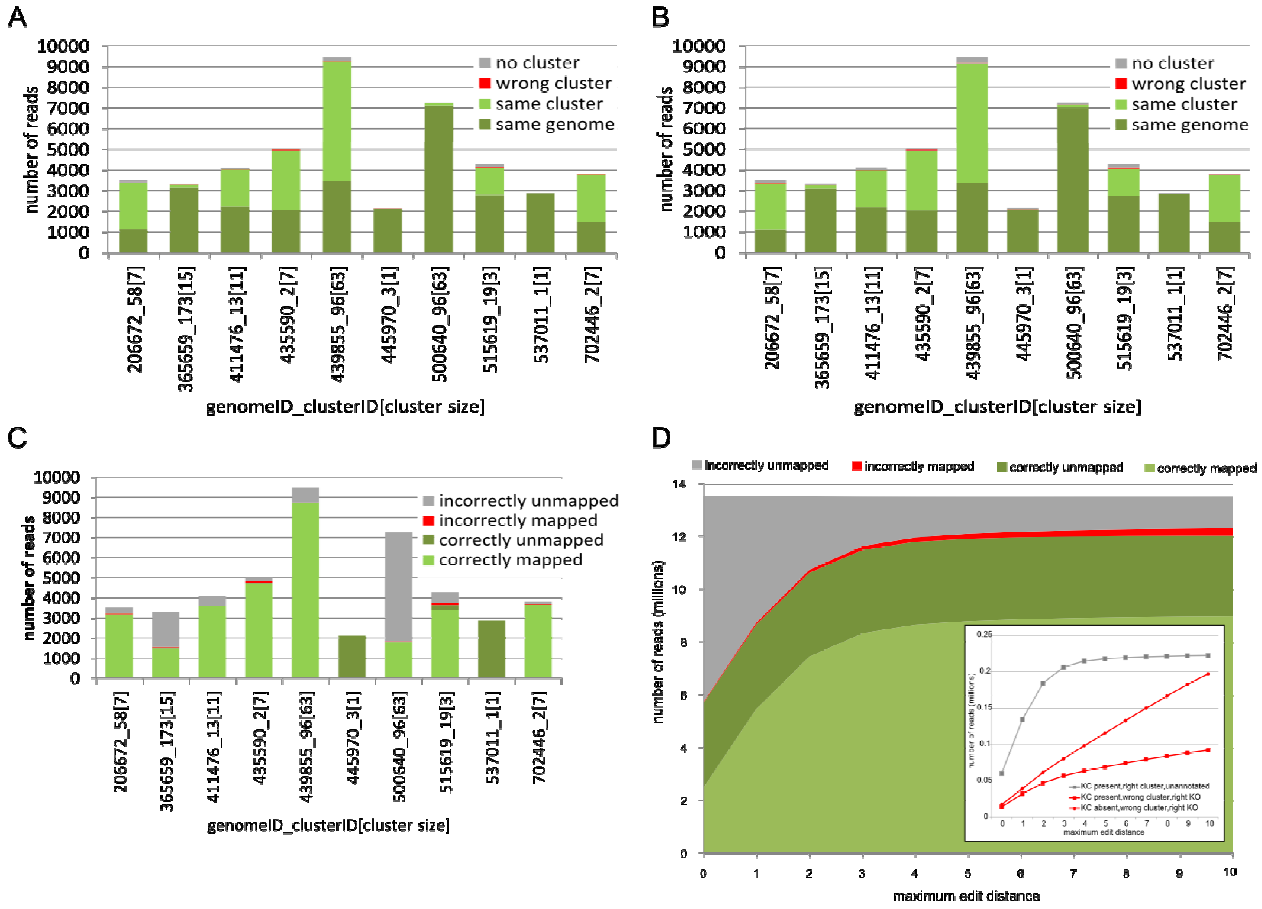


Figure C.1: Validation of reference genome clustering. 75 bp reads were simulated from 10 reference genomes, and then mapped back to a full or partial reference set using BWA (Extended Experimental Procedures). Each column represents the proportion of reads simulated from a single genome that fell into various mapping categories. In **(A)** reads were simulated from 10 selected genomes, and mapped to the full set of 260 reference genomes. In **(B)** reads were subject to a 1.5% sequencing-error model before being mapped to the reference genomes, and in **(C)** reads were subject to the error model and then mapped to a set of reference genomes in which the genome of origin was removed. As demonstrated, reads mapped successfully to the genome of origin or to an alternative genome in the same cluster if present, while very few reads mapped to the wrong cluster, supporting our cluster definitions. In **(D)** simulated mapping results as described in panel C were summed over all 260 genomes and analyzed at a range of maximum edit distances. An edit distance of 5 maximized the number of correctly mapped (or correctly unmapped) reads while minimizing the number of incorrectly mapped reads. The inset shows the most frequent assignments of incorrectly mapped reads for which the correct KC was present in the database (gray squares: unannotated regions in the correct genome cluster; red squares: correct KO in the wrong genome cluster) or absent from the database (red circles: correct KO in the wrong genome cluster).

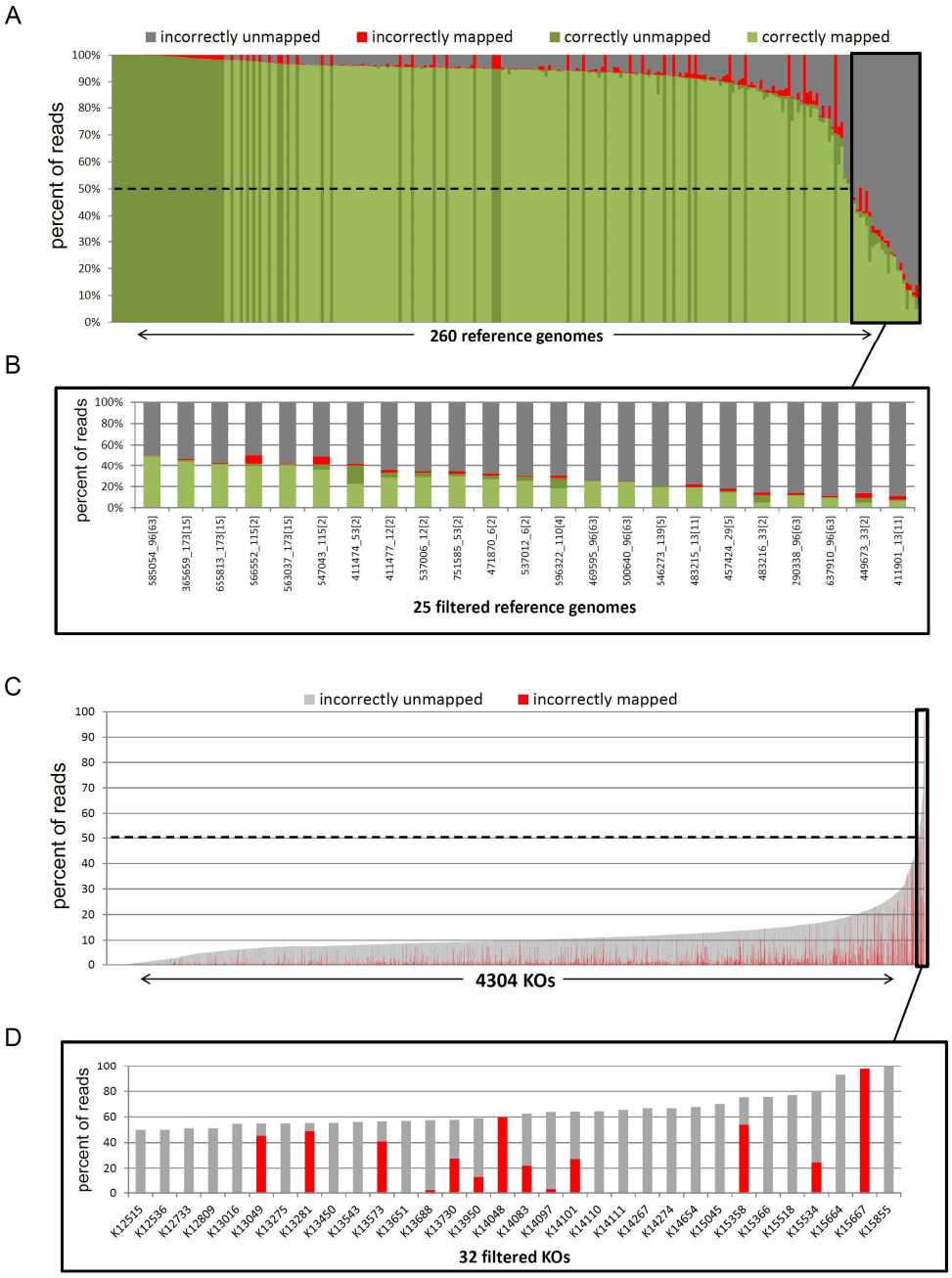
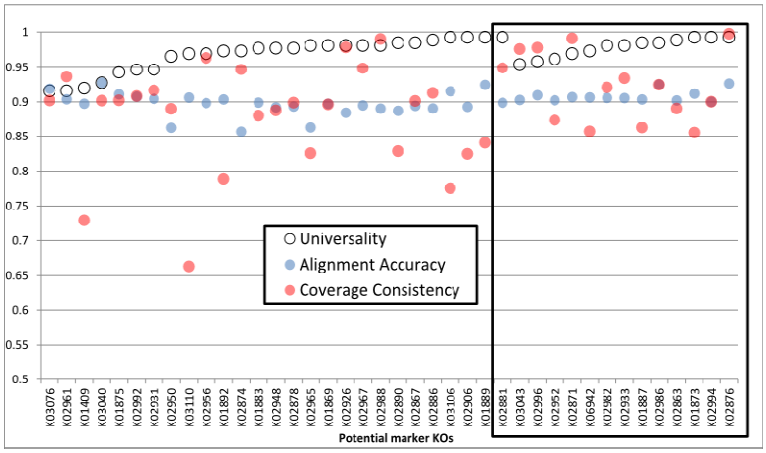


Figure C.2: Genomes and KOs with high mapping error rate. 75bp error-adjusted reads were mapped to a set of 260 reference genomes in which the genome of origin was removed (4.3 Methods; see also Fig. C.1C). The percent of reads simulated from each genome **(A)** or from each KO **(C)** that were correctly or incorrectly mapped are shown as stacked bars in each column. Genomes and KOs with a combined error rate (% incorrectly unmapped + % incorrectly mapped) >=50% were excluded from further analysis. The portions of panels A and C corresponding to these filtered genomes and KOs are magnified and shown in panels **(B)** and **(D)** respectively. Genome labels in panel B are formatted as *genomeID_clusterID[cluster size]* (and see Table S1).

A



B

KO ID	NAME	BRITE class
K03043	DNA-directed RNA polymerase, beta	Transcription machinery DNA repair and recombination
K02996	small subunit ribosomal protein S9	Ribosome
K02952	small subunit ribosomal protein S13	Ribosome
K02871	large subunit ribosomal protein L13	Ribosome
K06942	–	–
K02982	small subunit ribosomal protein S3	Ribosome
K02933	large subunit ribosomal protein L6	Ribosome
K01887	arginyl-tRNA synthetase	Amino acid related enzymes Transfer RNA biogenesis
K02986	small subunit ribosomal protein S4	Ribosome
K02863	large subunit ribosomal protein L1	Ribosome
K01873	valyl-tRNA synthetase	Amino acid related enzymes Transfer RNA biogenesis
K02994	small subunit ribosomal protein S8	Ribosome
K02876	large subunit ribosomal protein L15	Ribosome

Figure C.3: Selection of marker KOs. (A) 13 marker KOs were selected from a list of 40 potential KOs according to three criteria: Universality >0.95, Alignment Accuracy >0.90, and Coverage Consistency >0.85 and <1.15 (Extended Experimental Procedures). Selected marker KOs are outlined in black, and listed in **(B)**.

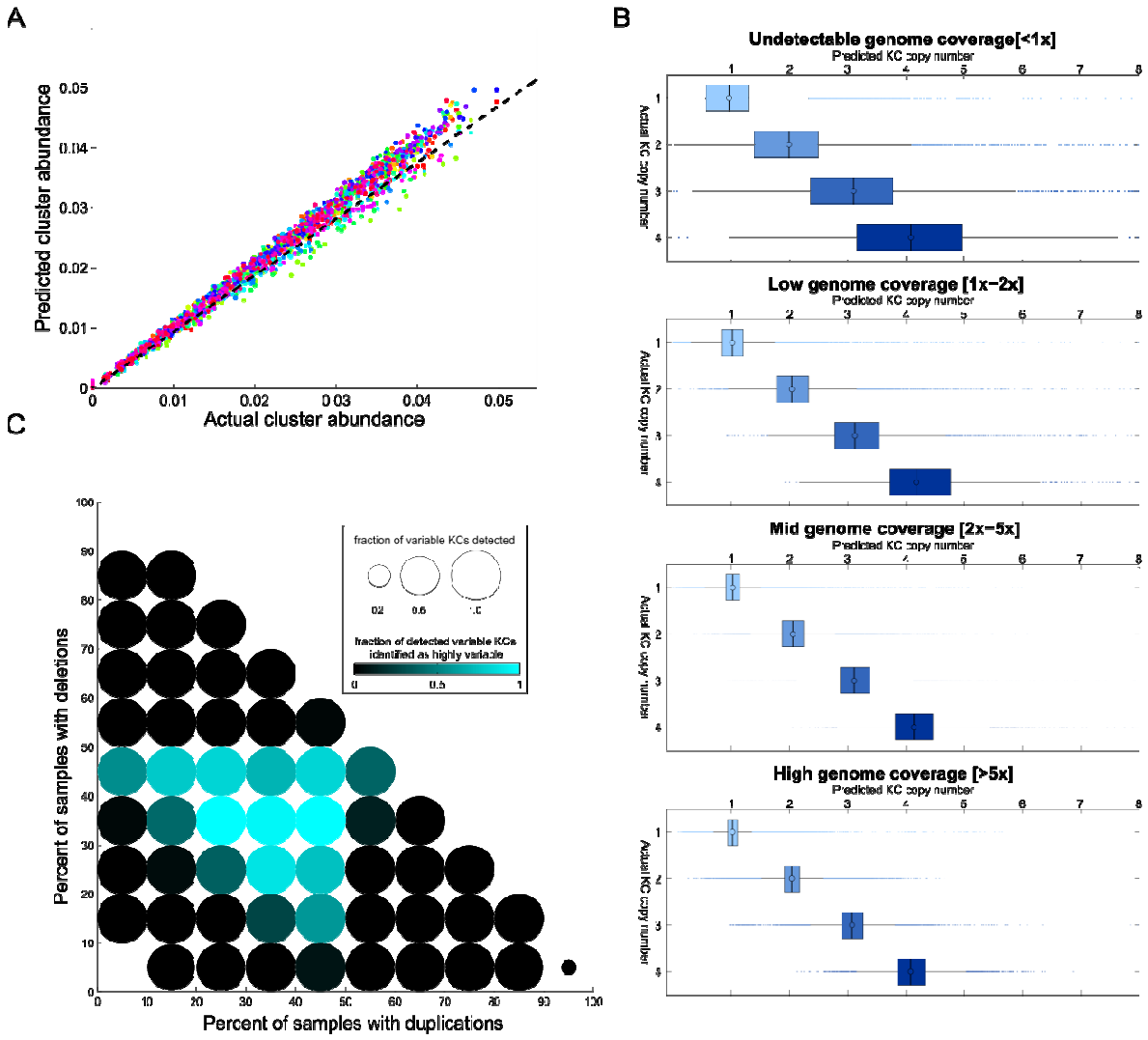


Figure C.4: Analysis of simulated reads from mock communities. (A) Predicted vs. real relative abundances of genome clusters across simulated communities (see C.1.7). Each point represents a single genome cluster in a single sample, with different clusters represented by different colors. Across all sampled and clusters, the correlation between predicted and real values was 0.993. **(B)** The estimated copy number of KCs as a function of the underlying real copy number and the coverage of the genome cluster. Each boxplots illustrates the distribution of copy number estimates obtained for KCs with a certain real copy number and in clusters with a given coverage range. Copy number estimation accuracy increased with coverage with an overall accuracy of 87.6%, 95.2%, and 97.8% for clusters with low coverage (1x-2x), intermediate coverage (2x-5x), and high coverage (>5x) respectively. Overall, copy number estimation accuracy for detectable clusters (>1x) was 96%, compared to only 60.1% for undetectable clusters. **(C)** Recall of variable KCs as a function of the fraction of samples in which the KC was deleted or duplicated. The color of each circle represents the proportion of these detected variable KCs that were also identified as highly variable (compared to set-specific variable), confirming the ability of our pipeline to classify the type of underlying variation.

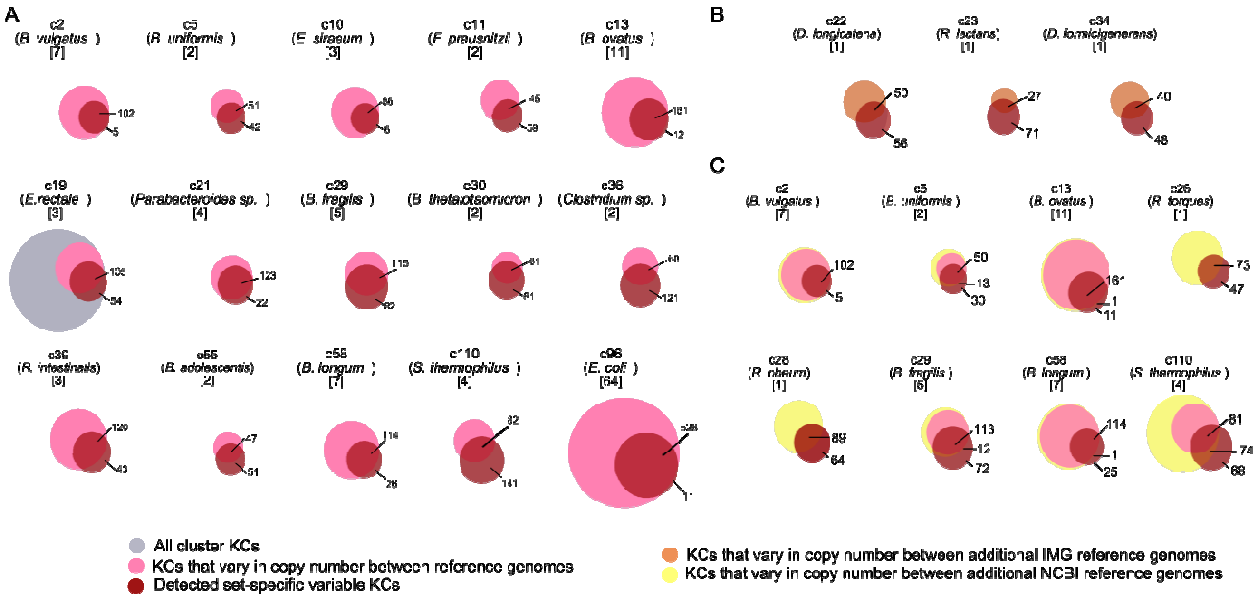


Figure C.5: Comparison of set-specific variable KCs to known variation among reference genomes. In each Venn diagram, the gray circle represents the set of all KCs in a given genome cluster, the pink circle represents the fraction of those KCs whose copy number varies across the cluster's reference genomes, and the red circle represents the set of set-specific variable KCs detected by our analysis. Overlap of the pink and red circles indicates correspondence between known and detected variation. Each diagram is labeled with the cluster ID, representative species name, and number of reference genomes. **(B-C)** Additional variation in reference genomes that were not used as mapping targets is represented by either an orange circle (additional reference genomes from IMG) or a yellow circle (additional reference genomes from NCBI), compared to variation in included reference genomes (pink) and detected set-specific variable KCs (red).

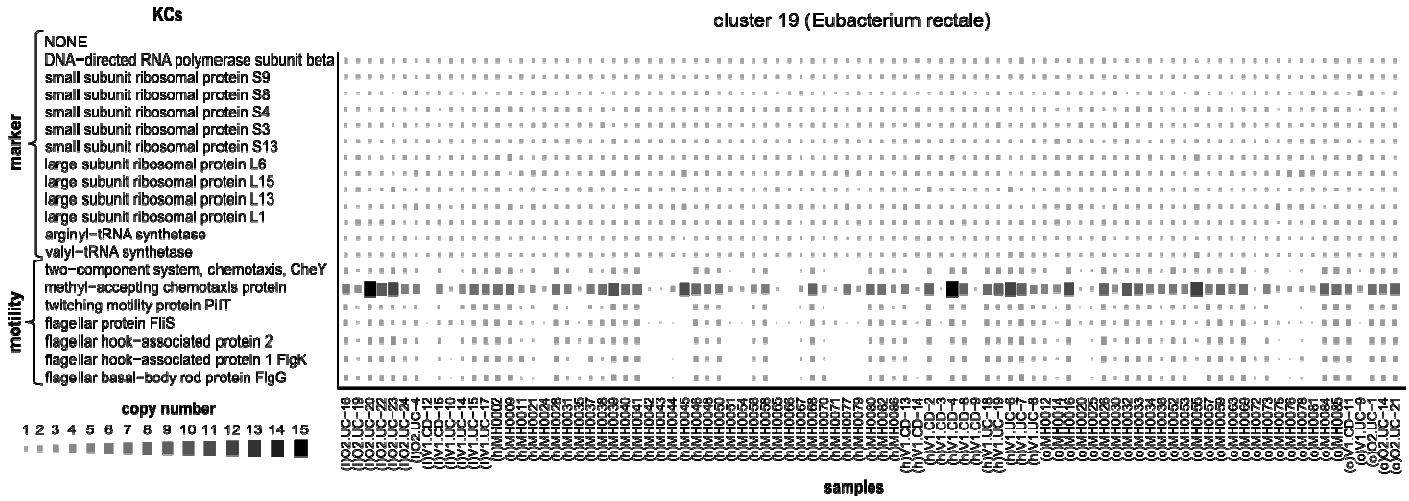


Figure C.6: Copy number of highly variable motility KCs in *Eubacterium rectale.* The size and color of each square represent the copy number of each highly variable KC within each sample. Samples are grouped by host state (I: IBD, h: healthy, o: obese). The copy number of the 13 marker KCs in this genome cluster and across the samples are also illustrated for comparison.

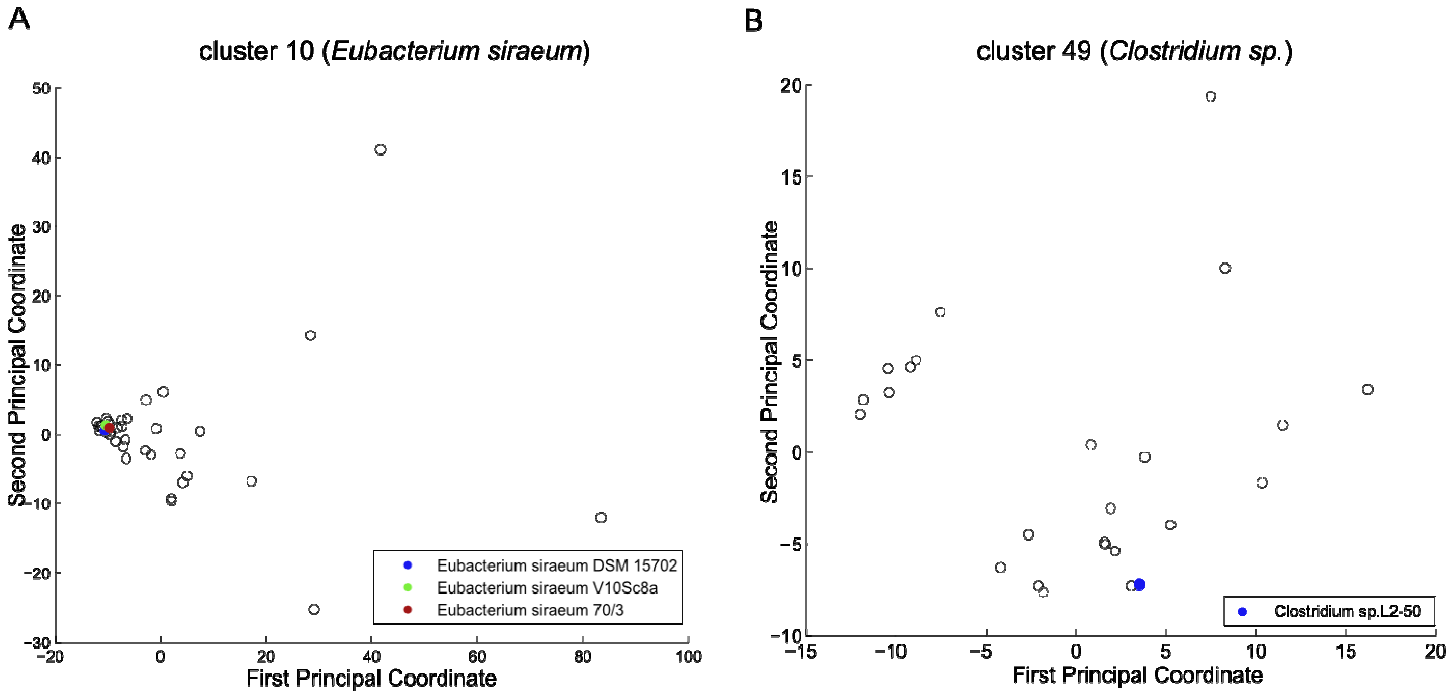


Figure C.7: Principal coordinate analysis of copy number profiles across samples. Principal coordinate plots are shown for two genome clusters: **(A)** *Eubacterium siraeum* and **(B)** *Clostridium sp.*, depicting differences between the copy number profiles across samples (open circles) and reference genomes (filled circles).

C.3 Supplementary Tables

Table C.1. Frequency of highly variable and set-specific variable KCs in each genome cluster.

Cluster	Representative Species	Number of Genomes	Sample count	Total KOs	Highly Variable KOs	Set-Specific Increased KOs	Set-Specific Decreased KOs	Set-Specific Inc/Dec KOs
1	Prevotella copri	1	26	868	4	39	45	7
2	Bacteroides vulgatus	7	99	1175	9	95	45	33
3	Alistipes putredinis	1	74	761	2	27	11	2
4	Ruminococcus bromii	1	51	699	4	27	12	8
5	Bacteroides uniformis	2	93	1075	7	53	60	20
7	Butyrivibrio crossotus	1	25	918	0	34	16	6
8	Bacteroides coprocola	1	21	942	9	47	34	15
9	Akkermansia muciniphila	1	26	879	1	23	21	5
10	Eubacterium siraeum	3	36	915	19	78	44	30
11	Faecalibacterium prausnitzii	2	73	949	25	102	46	45
13	Bacteroides ovatus	11	72	1205	34	143	109	79
14	Faecalibacterium prausnitzii	1	45	891	13	58	27	22
15	Ruminococcus sp.	1	61	1069	21	95	60	51
16	Eubacterium eligens	1	67	872	1	15	11	1
18	Faecalibacterium prausnitzii	1	76	870	25	107	47	47
19	Eubacterium rectale	3	91	1079	38	132	79	52
20	Roseburia inulinivorans	1	52	1045	44	135	77	68
21	Parabacteroides sp.	4	78	1138	25	122	69	46
22	Dorea longicatena	1	52	971	20	110	63	51
23	Ruminococcus lactaris	1	28	917	20	93	56	41
26	Ruminococcus torques	1	17	914	9	81	69	30
27	Coprococcus comes	1	48	946	19	76	53	38
28	Ruminococcus obeum	1	10	1017	38	132	72	51
29	Bacteroides fragilis	5	28	1145	22	115	140	58
30	Bacteroides thetaiotaomicron	2	47	1103	36	121	79	59
31	Dialister invisus	1	36	807	4	47	32	12
32	Bacteroides plebeius	1	25	981	7	121	23	9
34	Dorea formicigenerans	1	10	1001	38	86	73	45
35	Eubacterium ventriosum	1	12	891	19	71	43	29
36	Clostridium sp.	2	57	1026	15	113	120	52
37	Clostridium sp.	3	2	1107	0	0	0	0
38	Ruminococcus obeum	1	1	943	0	0	0	0
39	Roseburia intestinalis	3	40	1164	47	130	98	65
41	Bacteroides coprophilus	1	9	926	0	0	0	0
42	Coprococcus catus	1	2	938	0	0	0	0
43	Clostridium nexile	1	1	957	0	0	0	0
44	butyrate-producing bacterium	1	19	956	28	191	125	89
47	[Bacteroides] pectinophilus	1	8	919	0	0	0	0
48	Eubacterium hallii	1	29	1014	27	127	105	59
49	Clostridium sp.	1	24	919	17	56	142	27
51	Clostridium leptum	1	4	910	0	0	0	0
52	Lactobacillus reuteri	6	1	918	0	0	0	0
54	Acidaminococcus sp.	1	12	900	0	12	28	3

55	Bifidobacterium adolescentis	2	34	794	6	65	57	24
58	Bifidobacterium longum	7	16	941	11	109	58	27
62	Methanobrevibacter smithii	3	9	823	0	0	0	0
63	Ruminococcus gnavus	1	7	1036	0	0	0	0
64	Mitsuokella multacida	1	6	1081	0	0	0	0
66	Clostridium bolteae	1	3	1377	0	0	0	0
67	Clostridium hathewayi	1	2	1220	0	0	0	0
68	Catenibacterium mitsuokai	1	13	902	9	91	44	24
69	Blautia hansenii	1	3	976	0	0	0	0
72	Eubacterium biforme	1	4	828	0	0	0	0
88	Lactobacillus ruminis	1	4	889	0	0	0	0
89	Clostridium spiroforme	1	1	889	0	0	0	0
92	Desulfovibrio piger	1	9	1023	0	0	0	0
93	Clostridium bartlettii	1	2	1017	0	0	0	0
95	Ruminococcus sp.	1	2	749	0	0	0	0
96	Escherichia coli	63	13	3230	38	418	204	83
101	Acidaminococcus fermentans	1	2	945	0	0	0	0
102	Fusobacterium mortiferum	1	1	1036	0	0	0	0
110	Streptococcus thermophilus	4	11	991	24	133	128	38
131	Eggerthella lenta	1	1	933	0	0	0	0
139	Veillonella parvula	5	4	1018	0	0	0	0
161	Coprobacillus sp.	2	1	1058	0	0	0	0
180	Streptococcus parasanguinis	1	5	926	0	0	0	0
187	Bifidobacterium bifidum	1	4	780	0	0	0	0
213	Pediococcus pentosaceus	1	1	781	0	0	0	0
275	Collinsella stercoris	1	1	790	0	0	0	0
383	Klebsiella pneumoniae	6	2	2798	0	0	0	0

Table C.2: Functions enriched among highly and set-specific variable KCs.

Phylum	Genome Cluster	Annotation	Variation Type [Danish/ Spanish cohort]	Variation Type [Chinese cohort]*
Actinobacteria	*c55 (Bifidobacterium adolescentis)*	Putative multiple sugar transport system(M00207)	set-specific	n/a
		Putative spermidine/putrescine transport system(M00193)†	set-specific	
		Transporters(ko02000)†	set-specific	
Actinobacteria	*c58 (Bifidobacterium longum)*	Putative multiple sugar transport system(M00207)	high	n/a
Bacteroidetes	*c13 (Bacteroides ovatus)*	Galactose metabolism(ko00052)	set-specific	set-specific
		Iron complex transport system(M00240)†	high	high
		Lysosome(ko04142)†	set-specific	set-specific
		Other glycan degradation(ko00511)†	set-specific	set-specific
		Sphingolipid metabolism(ko00600)†	set-specific	set-specific
		Starch and sucrose metabolism(ko00500)	set-specific	
		Transporters(ko02000)†	high	high
Bacteroidetes	*c2 (Bacteroides vulgatus)*	Polyketide sugar unit biosynthesis (ko00523)†	set-specific	set-specific
Bacteroidetes	*c21 (Parabacteroides sp.)*	Amino sugar and nucleotide sugar metabolism(ko00520)†	set-specific	set-specific
Bacteroidetes	*c29 (Bacteroides fragilis)*	Amino sugar and nucleotide sugar metabolism(ko00520)†	set-specific	set-specific
		Cytoskeleton proteins(ko04812)†	high	
		DNA replication proteins(ko03032)†	set-specific	
		Iron complex transport system(M00240)†	high	
		Transporters(ko02000)	high	
Bacteroidetes	*c30 (Bacteroides thetaiotaomicron)*	Galactose metabolism(ko00052)	set-specific	set-specific
		Iron complex transport system(M00240)†	high	high
		Transporters(ko02000)	high	high
Bacteroidetes	*c32 (Bacteroides plebeius)*	Other glycan degradation(ko00511)†	set-specific	set-specific
Firmicutes	*c11 (Faecalibacterium prausnitzii)*	Riboflavin biosynthesis, GTP => riboflavin/FMN/FAD(M00125)†	set-specific	high
		Riboflavin metabolism(ko00740)†	set-specific	high
Firmicutes	*c110 (Streptococcus thermophilus)*	Ribosome(ko03010)†	set-specific	n/a
		Ribosome(ko03011)†	set-specific	
		Ribosome, bacteria(M00178)†	set-specific	
		Transporters(ko02000)	set-specific	
Firmicutes	*c14 (Faecalibacterium prausnitzii)*	Iron complex transport system(M00240)†	set-specific	
		PTS system, mannose-specific II component(M00276)†	set-specific	
		Putative multiple sugar transport	set-specific	

		system(M00207)†		
		Transporters(ko02000)†	set-specific	high
Firmicutes	*c15 (Ruminococcus sp.)*	Putative multiple sugar transport system(M00207)	high	high
		Transporters(ko02000)†	high	high
Firmicutes	*c18 (Faecalibacterium prausnitzii)*	Oligopeptide transport system(M00439)†	set-specific	set-specific
		Transporters(ko02000)†	high	high
Firmicutes	*c19 (Eubacterium rectale)*	Bacterial motility proteins(ko02035)†	high	
		Putative multiple sugar transport system(M00207)	high	
		Transporters(ko02000)†	high	high
Firmicutes	*c20 (Roseburia inulinivorans)*	ATP synthase(M00164)†	set-specific	set-specific
		F-type ATPase, bacteria(M00157)†	set-specific	set-specific
		Photosynthesis proteins(ko00194)†	set-specific	set-specific
		Photosynthesis(ko00195)†	set-specific	set-specific
		Transporters(ko02000)†	high	high
Firmicutes	*c22 (Dorea longicatena)*	ABC transporters(ko02010)†	set-specific	n/a
		Transporters(ko02000)†	high	
Firmicutes	*c23 (Ruminococcus lactaris)*	Fructose and mannose metabolism(ko00051)†	set-specific	
Firmicutes	*c26 (Ruminococcus torques)*	Peptidases(ko01002)†	set-specific	
Firmicutes	*c27 (Coprococcus comes)*	ABC transporters(ko02010)†	set-specific	n/a
		Transporters(ko02000)†	set-specific	
Firmicutes	*c28 (Ruminococcus obeum)*	Transcription factors(ko03000)†	set-specific	n/a
		Transporters(ko02000)†	set-specific	
Firmicutes	*c31 (Dialister invisus)*	ABC transporters(ko02010)†	set-specific	n/a
		ADP-L-glycero-D-manno-heptose biosynthesis(M00064)†	set-specific	
		Iron complex transport system(M00240)†	set-specific	
		Iron(III) transport system(M00190)†	set-specific	
		Lipopolysaccharide biosynthesis proteins(ko01005)†	set-specific	
		Nickel transport system(M00440)†	set-specific	
		Peptides/nickel transport system (M00239)†	set-specific	
		Transporters(ko02000)†	set-specific	
Firmicutes	*c34 (Dorea formicigenerans)*	Oligopeptide transport system(M00439)†	set-specific	n/a
		Transporters(ko02000)	set-specific	
Firmicutes	*c35 (Eubacterium ventriosum)*	Transporters(ko02000)	set-specific	set-specific
Firmicutes	*c36 (Clostridium sp.)*	Pentose phosphate pathway(ko00030)†	set-specific	
		Phosphotransferase system (PTS) (ko02060)†	set-specific	set-specific

		Transporters(ko02000)†	set-specific	set-specific
Firmicutes	*c39 (Roseburia intestinalis)*	Transporters(ko02000)	set-specific	high
Firmicutes	*c44 (butyrate-producing bacterium)*	Bacterial chemotaxis(ko02030)†	set-specific	
		Bacterial motility proteins(ko02035)†	set-specific	set-specific
		CheA-CheYBV (chemotaxis) two-component regulatory system(M00506)†	set-specific	
		Flagellar assembly(ko02040)†	set-specific	set-specific
		Peptides/nickel transport system (M00239)†	set-specific	high
		Ribosome(ko03011)†	set-specific	set-specific
		Secretion system(ko02044)†	set-specific	set-specific
Firmicutes	*c49 (Clostridium sp.)*	Bacterial motility proteins(ko02035)†	set-specific	
		Chromosome(ko03036)†	set-specific	
		Cytoskeleton proteins(ko04812)†	set-specific	
		Lipopolysaccharide biosynthesis proteins(ko01005)†	set-specific	
		Peptidases(ko01002)†	set-specific	
		Secretion system(ko02044)†	set-specific	
Firmicutes	*c54 (Acidaminococcus sp.)*	Polyketide sugar unit biosynthesis (ko00523)†	set-specific	n/a
		Streptomycin biosynthesis(ko00521)†	set-specific	
Firmicutes	*c68 (Catenibacterium mitsuokai)*	Fructose and mannose metabolism(ko00051)†	set-specific	n/a
		Phosphotransferase system (PTS) (ko02060)†	set-specific	
		Transporters(ko02000)†	set-specific	
Firmicutes	*c7 (Butyrivibrio crossotus)*	Putative ABC transport system(M00258)	set-specific	n/a
Proteobacteria	*c96 (Escherichia coli)*	Bacterial secretion system(ko03070)†	set-specific	set-specific
		D-Allose transport system(M00217)†	set-specific	
		EHEC/EPEC pathogenicity signature, T3SS and effectors(M00542)†	set-specific	set-specific
		Fructose and mannose metabolism(ko00051)†	set-specific	
		Iron complex transport system(M00240)	high	high
		Manganese/iron transport system (M00317)†	set-specific	
		Pertussis(ko05133)†	set-specific	
		Phenylalanine metabolism(ko00360)†	set-specific	
		Phosphotransferase system (PTS) (ko02060)†	set-specific	set-specific
		PTS system, fructose-specific II component(M00273)†	set-specific	
		PTS system, galactitol-specific II	set-specific	

		component(M00279)†		
		PTS system, sorbose-specific II component(M00278)†	set-specific	
		Ribose transport system(M00212)†	set-specific	set-specific
		Secretion system(ko02044)†	set-specific	set-specific
		Transcription factors(ko03000)†	set-specific	set-specific
		Transporters(ko02000)†	set-specific	
		Type II general secretion system(M00331)†	set-specific	set-specific
		Type III secretion system(M00332)†	set-specific	set-specific
Verrucomicrobia	*c9 (Akkermansia muciniphila)*	Other glycan degradation(ko00511)	set-specific	n/a
		Sulfur metabolism(ko00920)†	set-specific	
		Sulfur reduction,sulfate => H2S(M00176)†	set-specific	

† Enrichment robust to filtering KCs with high median copy number (Extended Experimental Procedures).

* Genome clusters marked as n/a were not present in a sufficient number of samples in the Chinese cohort to infer copy number variation.

Table C.3. Host-state associated copy number variation.

Genome cluster	KO	KO name	KEGG PW	p-val	Inc/Dec
IBD-associated KCs					
c55 (Bifidobacterium adolescentis)	K01624	fructose-bisphosphate aldolase, class II [EC:4.1.2.13]	Glycolysis / Gluconeogenesis (+many others)	1.12E-04	dec
c55 (Bifidobacterium adolescentis)	K01421	putative membrane protein	none	2.44E-04	dec
c20 (Roseburia inulinivorans)	K08217	MFS transporter, DHA3 family, macrolide efflux protein	Transporters	4.07E-08	inc
c2 (Bacteroides vulgatus)	K01153	type I restriction enzyme, R subunit [EC:3.1.21.3]	Hydrolases	9.92E-05	inc
c20 (Roseburia inulinivorans)	K05593	aminoglycoside 6-adenylyltransferase [EC:2.7.7.-]	Transferases	1.29E-04	inc
c2 (Bacteroides vulgatus)	K05593	aminoglycoside 6-adenylyltransferase [EC:2.7.7.-]	Transferases	1.38E-04	inc
c14 (Faecalibacterium prausnitzii)	K01246	DNA-3-methyladenine glycosylase I [EC:3.2.2.20]	Base excision repair	1.74E-04	inc
c2 (Bacteroides vulgatus)	K01154	type I restriction enzyme, S subunit [EC:3.1.21.3]	Hydrolases	2.21E-04	inc
c22 (Dorea longicatena)	K05593	aminoglycoside 6-adenylyltransferase [EC:2.7.7.-]	Transferases	2.98E-04	inc
c5 (Bacteroides uniformis)	K03296	hydrophobic/amphiphilic exporter-1 (mainly G- bacteria), HAE1 family	none	4.25E-04	inc
c2 (Bacteroides vulgatus)	K01919	glutamate--cysteine ligase [EC:6.3.2.2]	Glutathione metabolism	8.42E-04	inc
c5 (Bacteroides uniformis)	K09686	antibiotic transport system permease protein	ABC transporters	8.64E-04	inc
c5 (Bacteroides uniformis)	K01993	HlyD family secretion protein	none	9.89E-04	inc
c5 (Bacteroides uniformis)	K01990	ABC-2 type transport system ATP-binding protein	Transporters	1.22E-03	inc
c16 (Eubacterium eligens)	K02114	F-type H+-transporting ATPase subunit epsilon [EC:3.6.3.14]	Oxidative phosphorylation|Photosynthesis	1.79E-03	inc
c2 (Bacteroides vulgatus)	K02003	putative ABC transport system ATP-binding protein	Transporters	2.06E-03	inc
c2 (Bacteroides vulgatus)	K03111	single-strand DNA-binding protein	DNA replication|Mismatch repair|Homologous recombination	2.43E-03	inc
c5 (Bacteroides uniformis)	K16089	outer membrane receptor for ferrienterochelin and colicins	Transporters	2.62E-03	inc
c2 (Bacteroides vulgatus)	K00561	23S rRNA (adenine2085-N6)-dimethyltransferase [EC:2.1.1.184]	Ribosome biogenesis,	2.65E-03	inc
c2 (Bacteroides vulgatus)	K01738	cysteine synthase A [EC:2.5.1.47]	Cysteine and methionine metabolism|Sulfur metabolism	3.01E-03	inc
c5 (Bacteroides uniformis)	K06180	23S rRNA pseudouridine1911/1915/1917 synthase [EC:5.4.99.23]	Ribosome biogenesis	3.07E-03	inc
c5 (Bacteroides uniformis)	K01686	mannonate dehydratase [EC:4.2.1.8]	Pentose and glucuronate interconversions	3.95E-03	inc
c5 (Bacteroides uniformis)	K01897	long-chain acyl-CoA synthetase [EC:6.2.1.3]	Fatty acid metabolism|PPAR signaling pathway|Peroxisome|	4.12E-03	inc
c5 (Bacteroides uniformis)	K03924	MoxR-like ATPase [EC:3.6.3.-]	Hydrolases	4.83E-03	inc
Obesity-associated KCs					
c32 (Bacteroides	K00349	Na+-transporting NADH:ubiquinone	Oxidoreductases	3.51E-04	dec

plebeius)		oxidoreductase subunit D [EC:1.6.5.-]			
c49 (Clostridium sp.)	K03671	thioredoxin 1	Protein folding catalysts	1.13E-04	inc
c49 (Clostridium sp.)	K01153	type I restriction enzyme, R subunit [EC:3.1.21.3]	Hydrolases	1.38E-04	inc