Assembly Rules of the Microbiome

Roie Levy

A dissertation submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

University of Washington

2015

Reading Committee:

Elhanan Borenstein, Chair
Herbert Sauro
Benjamin Kerr

Program authorized to offer degree: Molecular & Cellular Biology

University of Washington

Abstract

Assembly rules of the Microbiome

Roie Levy

Supervisory committee chair:
Elhanan Borenstein, Associate Professor
Department of Genome Sciences

The microbiome – the communal ecosystem of bacteria, archaea, and microscopic eukarya as well as the environments sustaining them – surpasses its macroscopic counterpart in taxonomic diversity, metabolic capabilities, and biogeographical distribution. Microbes inhabit every habitat hospitable to life, even those too extreme for higher order organisms. In milder environments, microbes provide the foundation for all life, either by maintaining essential biogeochemical cycles or by directly supporting host macroorganism health. These microbes interact to form complex communities, some of which contain up to thousands of distinct species. Modern metagenomic techniques have revealed extensive structure in these communities in the form of non-random species co-occurrence patterns. It has been proposed that patterns such as these arise through cogent *assembly rules* – deterministic ecological processes that govern the characteristic structure of a community. Yet, because the physiology and behavior of only a tiny fraction of microorganisms has been characterized, it becomes challenging to distinguish between alternative sets of community assembly rules. In this

dissertation, I describe work I performed during my doctoral studies to develop analytical frameworks that integrate community metagenome information with single species whole genome information to identify community assembly rules that structure the human and global microbiomes. In chapter 1 I discuss the context for this work: the microbial communities studied, the methods used to characterize them, and the specific challenges I aim to address. In chapter 2 I describe how I used metabolic models of species interaction to determine that habitat filtering, and not species assortment, determines the structure of the human microbiome. These models additionally showed that host intestinal health state does not define the axes along which the community is filtered, hinting at more subtle biochemical processes yet to be determined. In chapter 3 I describe an analysis of functional complementarity, interactions where microbes bring to a shared environment a pair of functions not typically encoded within a single genome. Investigating this interaction on a large scale across a network of co-occurring microbes, I demonstrate that niche partitioning, rather than cooperative interaction, structures the global microbiome. In chapter 4 I describe a comparison of genetic co-occurrence across genomes and metagenomes. Differential analysis of co-occurrence structure reveals that while metagenome structure resembles that found in genomes, processes such as environmental remediation are still likely to be distributed across community members. Finally, in chapter 5 I discuss how these particular assembly rules relate to one another at different ecological scales, and offer perspective as to the future development and application of the analytical frameworks presented here.

Acknowledgments

I cannot express sufficiently my gratitude for my adviser, Elhanan, a true mentor. From our first conversation it was clear that he would be source of scientific inspiration; what continues to surprise is his dedication to my own growth not only scientifically, but academically, professionally, and personally. I could hope for no better a companion with whom to formulate, investigate, and occasionally answer the countless offbeat questions I asked during my studies.

I dedicate this work to the memory of Stuart Davidson, a colleague, a friend, and a partner in crime. His enthusiasm in all things was infectious, and he had limitless devotion to his friends. He was always the first to celebrate my brightest moments, and during the darkest he managed to accomplish what no else had: he convinced me to take care of myself. To say that I would not have made it where I am without him may be cliché, but it is no overstatement.

Contents

Table Of Figures

1. Introduction

Microbial ecology, the study of microbial biogeography, communities, and interactions, has seen a surge in interest from the scientific community over recent decades. Fueled largely by technological innovations in the field of gene and genome sequencing, researchers have not only made great strides in cataloguing diversity along the entire microbial tree of life, but have begun to investigate the myriad communities inhabiting the human body, the world's oceans and soils, and even extreme environments inhospitable to all but the most specialized life forms. As these communities' compositions were characterized, the intrinsic complexity of these microscopic ecosystems – termed *microbiomes* – became unmistakable. In the most extreme cases, a microbiome may contain thousands of species, all interacting with one another as well as with an unpredictable environment. Consequently, parallels between macroscopic ecology and the ecology of microorganisms became apparent, and longstanding questions in classical ecology resurfaced in the microbial world. In particular the existence of community assembly rules, deterministic ecological processes which act on communities to provide their characteristic structure, came under debate. Because physiological characterization of constituent microbes was lacking, distinguishing between alternative modes of community assembly proved difficult. Ultimately, novel analyses combining systems approaches, whole genome modeling, and a community ecology framework would become necessary to address this challenge.

1.1. The importance of microbial communities, and challenges to their study

The importance of the microbiome cannot be understated, yet it is often overlooked. Microbes inhabit any environment capable of supporting life, and are responsible for providing the foundation necessary for more complex organisms. Free living microbes are major contributors to biogeochemical processes across the globe, while endosymbionts contribute essential functionality to their hosts. In many cases microbial consortia are responsible for these phenomena; independently acting species are incapable of or ineffective at coordinating the necessary processes. Owing to its critical medical relevance, the human microbiome in particular has received much attention, but on a global scale a diversity of ecologically relevant microbiomes contribute to crop productivity, environmental remediation, and overall ecosystem health.

1.1.1. The human microbiome

The human body is home to trillions of microbial organisms. By some estimates, the number of microbial cells outnumbers host cells 10:1 [1]. This community is exceptionally diverse; recent surveys place the number of species in the thousands [2]. The genetic diversity of this microbiome surpasses its taxonomic diversity. The *metagenome* – the collection of genes and genomes belonging to the organisms in these communities – contains within it 150 times as many genes as does the human genome [3]. In many cases, these genes encode functions which the host is incapable of performing intrinsically, and for which it relies on its microbial endosymbionts. These functions include, for example, the processing of complex

polysaccharides [4–6], the breakdown of toxic compounds [7–10], resistance to pathogens [11,12], and informing innate immunity [13].

Importantly, in all but a few cases, a single species has not been identified as the sole contributor of these host-critical functions. While in some cases this might simply reflect redundancy in the capabilities of community members, this may also reflect an interaction between species or an emergent property of community diversity as a whole. It has been shown, for example, that chemotrophic archaea sometimes found in the human gut feed off the byproducts of sulfate reducing bacteria, producing methane as an end product [14,15]. Metabolic interactions among gut microbes form a complex network in which degraders of complex polysaccharides make available simpler sugars which are subsequently fermented by other microbes [16]. These interactions are neither simple nor straightforward; gut microbes have been shown to sense and respond to their neighbors by diversifying metabolic strategies, yet at the same time exchange metabolic intermediates [17]. Such complex interactions, coupled to the inherently convoluted nature of the community, complicates the determination of the specific role the community, as well as its constituents, plays in influencing the health of the human host.

For this reason it becomes challenging to distinguish to what extent the microbiome drives host health, as opposed to responding to host condition. In many cases the association between the microbiome and host health is supported only by differential analysis of community composition. Such approaches not only cannot distinguish cause from effect, they also do not provide a mechanistic understanding of the drivers of disease; typically they only implicate specific taxonomic or gene families. Also, because they lack appropriate physiological context,

statistical analyses may be confounded by uncontrolled extrinsic factors. This may be why, for example, while the ratio of *firmicutes* to *bacteroidetes* has been implicated in obesity, there is no consensus as to which phylum dominates in health versus obesity [18,19]. Experiments in gnotobiotic mouse model systems have been used extensively to test hypothesized modes of action, but these too may not directly identify specific molecular mechanisms or the set of species responsible [20].

1.1.2. The global microbiome

Microbes are found in every habitat capable of supporting life. Even seemingly inhospitable environments like hypersaline lakes and acid mine drainages can support communities with on the order of 10 species [21,22], while richer environments support communities with orders of magnitudes greater richness. Seawater communities comprise approximately 150 species [23], while a single gram of soil contains 4,600 to 10,000 distinct species [24,25]. Global microbial biogeography indicates that many of these species have highly specific co-occurrence patterns, possibly indicating co-mingled lifestyles or interactions essential for survival [26]. In many instances, interactions between these co-occurring species drive biogeochemical processes of great ecological importance.

These interactions often occur between organisms remarkably distant from one another on the tree of life. In aquatic sediment environments, a consortium of sulfate reducing bacteria (SRB) and anaerobic methanotrophic archaea (ANME) is responsible for oxidizing methane, a greenhouse gas [27]. The blue-green algae of the phylum cyanobacteria that are responsible for the initial, and largely the continued, oxygenation of Earth's atmosphere are known to interact with an assortment of microbes [28,29]. Microbial communities in the rhizosphere are critical to

the health and productivity of indigenous plants, in part through their role in making bioavailable nitrogen. Notably, community diversity is a critical contributor to nitrogen fixation efficiency [30]. Thus interactions between microbes contribute not only to their own survival, but to the well-being of all life on Earth.

1.2. Specific methods of studying the microbiome

Advanced genome sequencing technologies are credited with catalyzing the recent wave of research in microbial ecology, but DNA sequencing has always been a cornerstone in the study of these communities. As many as 99% of species in any environmental sample resist culture and isolation, in effect making molecular techniques essential to the taxonomic characterization microbes and their communities [31]. It was only through analysis of non-coding gene sequences that the structure of the tree of life could be characterized [32]. Currently, culture-free molecular investigation of microbial communities is largely dominated by two paradigms: targeted and shotgun metagenomics. Broadly, the former focuses on identifying the species present in an environmental sample, while the latter seeks to identify the functional capacity of the community *en masse*.

1.2.1. Targeted metagenomics

Targeted metagenomic surveys amplify and analyze marker sequences in order to assign a taxonomic profile to environmental samples. Typically, the 16S non-coding ribosomal gene is targeted, as it has been established as a universal phylogenetic marker gene present in all prokaryotic genomes. It contains within its sequence conserved as well as hypervariable regions, allowing PCR or sequencing primers to be designed which target those areas with the greatest phylogenetic information [33]. Subsequently, bulk 16S DNA can be extracted from the

environment and quantified without an intermediate culturing step, providing the relative abundance of each phylotype [34].

Ultimately, this methodology provided a relatively cheap means to characterize microbial biogeography on a large scale. In fact, as sequencing technologies improved, phylogenetic profiles of thousands of environments were generated. For the first time, microbial ecologists had sufficient species-incidence information to perform large scale analyses of the global microbiome, as well as particular microbiomes of interest. Non-random co-occurrence patterns revealed specific associations between species across both the human and global microbiomes [26,35]. Such patterns have been taken as evidence of ecological structure, leading to the first hypotheses that microbial communities may be structured by similar niche processes as their macroecological counterparts [36,37].

1.2.2. Shotgun metagenomics

More recently, shotgun metagenomics has become an alternative to the targeted approach. Using this method DNA is extracted and analyzed directly from the environment as in targeted metagenomics, but any DNA from any microbial genome present may be sequenced. Depending on the research objective, the sequencing platform employed, the depth of sequencing, and the complexity of the community, different steps may be taken to characterize the community. In the simplest approach, individual reads are annotated by homology search against a database of genes with known function. Most often, adjoining reads are tiled into longer contiguous assemblies, along which whole genes can be identified. Regardless of methodology, the number of reads that map to genes represents not only functional capacity of

the community, but can reveal which functions are enriched or depleted among community members.

Shotgun metagenomics effectively determines the structure of communities at the functional level. Using this approach, entire genomes from a low-complexity acid mine drainage community were reconstructed to near completion [22]. Such an approach indicated potential metabolic interactions between community members as well as with the environment. Shotgun metagenomics of the healthy human microbiome showed that while communities vary immensely in taxonomic representation, the functional profiles may be more constrained [2]. This potentially indicates selection acting on the community for a particular suite of functions, irrespective of which microbes perform which functions. In this manner, both targeted and shotgun metagenomic approaches have revealed strong structural signatures in microbial communities.

1.3. Metagenomics to address questions in microbial ecology

Metagenomic characterization of microbial communities may have identified structural patterns, but analysis of those patterns was found to be insufficient for unambiguously determining underlying structuring principles. Initial attempts to do so typically applied to the microbiome analytical techniques established in the field of classical ecology. These analyses primarily concerned questions of taxonomic diversity, productivity, and response to perturbation (including shifts in host health). It became clear that communities contained significantly different numbers of species across as well as within environments [25,38], the distributions of which followed patterns observed in macroscopic communities [36]. The organization of macroscopic communities, however, is relatively easy to characterize. The physiology of plants

and animals has been studied for centuries, and interactions such as predation are easily observed. A mechanistic model of these interactions, in turn, was critical in informing models of macroscopic community assembly.

In comparison to plants and animals, almost nothing is known of the physiology of most microbes. The utilization of sequence information not only as a means to characterize community composition, but to provide identity to constituent species, provides a potential solution to this challenge. 16S sequence, for example, not only provides a measure of an organism's abundance but reveals an organism's taxonomic background. By integrating the two, phylogenetic analysis of rank-abundance distributions can quantify the contribution of niche and neutral processes in gut communities [37]. Similar analyses demonstrate phylogenetically clustering in environmental microbiomes, implicating habitat filtering as a structuring force [39]. Nonetheless, because phylogenetic information is only a proxy of physiology, it does not provide a mechanistic model of the community assembly process.

1.4. Integrating metagenomics and genome scale models to identify community assembly rules

To address these open questions in microbiome assembly, it becomes necessary to integrate whole-genome analysis of individual organisms with metagenomic-derived community composition. In doing so, each constituent member of the community not only receives a phylogenetic identity, but a physiological characterization. These in turn inform mechanistic models of species interaction, which can be associated with specific models of community assembly *a priori*. The chapters that follow describe work I performed to develop analytic frameworks which integrate whole genome and metagenome information to answer three

specific questions regarding microbiome assembly. Specifically, these are (1) is the human microbiome structured by habitat filtering or by species assortment? (2) Is the global microbiome structured by cooperative interactions more so than by niche partitioning? And (3) what information regarding community assembly can be inferred from structure found in a metagenome that is not found in prokaryotic genomes?

1.4.1. Habitat filtering versus species assortment

Microbial communities, like their macroscopic counterparts, exhibit a significant number of checkerboards: pairs of species which exclude one another more than expected by chance. In order to explain this phenomenon, Diamond defined the concept of community assembly rules, and in the process described the species assortment model: competitive interactions drive these species apart [40]. Soon after, Connor and Simberloff proposed the habitat filtering model as an alternative: species occupy non-overlapping niches available in only a subset of habitats, and in turn segregate across environments [41]. To distinguish between these models, I used genome-scale metabolic models to predict metabolic interactions between human-associated microbes. Under the habitat filtering model, species would co-occur most with their competitors; under the species assortment model, species would segregate from their competitors.

1.4.2. Niche partitioning versus distributed metabolism

Microbes are capable of syntrophy, a cooperative interaction in which one organism feeds off the metabolic byproducts of another. It has been proposed that microbes employ a strategy of distributed metabolism whereby auxotrophic bacteria complement one another's metabolic needs [42]. Conversely, microbes also sense the presence of potential competitors among their neighbors, and respond by altering their metabolic programs in order to focus on

non-overlapping set of nutrients, a process known as niche partitioning [17]. Both processes account for non-random co-occurrence profiles: the former by obligate mutualism, the second by competition mitigation. A signature of both processes is the juxtaposition of different functional capabilities by distinct species. To determine which strategy is a greater contributor to global microbiome co-occurrence structure, I performed a large scale analysis of functional complementarity in the genomes of co-occurring bacteria.

1.4.3. Assembly of the community metagenome

Similar to their biogeography, microbes' genomes exhibit structure in the form of non-random genetic co-occurrence [43]. These structural patterns reflect functional associations as well as vertical co-inheritance [44]. Because a metagenome is a weighted combination of individual genomes, co-occurrence structure of the metagenome may be expected to resemble structure found in individual genomes. Nonetheless, because community assembly is a stochastic process, one might expect noisy admixture of genomes to obscure this structure. Additionally, because the relative abundance of a genome varies in accordance with its fitness across environments, significant genetic co-occurrences may be found across metagenomes not found across genomes. To identify associations among genes driven by environmental interactions rather than direct interactions or vertical co-inheritance, I performed differential analysis of genetic co-occurrence in metagenomes and genomes.

2. Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules

This chapter is based on the manuscript:

Levy R, Borenstein S. (2013) Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules. Proceeding of the National Academy of Sciences of the United States of America 110(31), 12804–12809.

This was subsequently expanded upon in another manuscript:

Levy R, Borenstein E. (2014) Metagenomic systems biology and metabolic modeling of the human microbiome. Gut Microbes (2014) 5:2, 1–6.

2.1. Summary

The human microbiome plays a key role in human health and is associated with numerous diseases. Metagenomics studies generate valuable information about the composition of the microbiome in health and in disease, and demonstrate non-neutral assembly processes and complex co-occurrence patterns. However, the underlying ecological forces that structure the microbiome are still unclear. Specifically, compositional studies alone, with no information about mechanisms of interaction such as competition or syntrophy, cannot clearly distinguish habitat-filtering and species assortment assembly processes. To address this

challenge, I introduce a computational framework that integrates metagenome-based compositional data with genome-scale metabolic models of species interaction. I use *in silico* metabolic network models to predict levels of competition and complementarity among 154 microbiome species and compare predicted interaction measures to species co-occurrence. Applying this approach to two large-scale datasets describing the composition of the gut microbiome, I find that species tend to co-occur across individuals with species with which they strongly compete, suggesting that microbiome assembly is dominated by habitat filtering. Moreover, species' partners and excluders exhibit distinct metabolic interaction levels. Importantly, I show that these trends cannot be explained by phylogeny alone and hold across multiple taxonomic levels. Interestingly, controlling for host health does not change the observed patterns, indicating that the axes along which species are filtered are not fully defined by macroecological host states. The approach presented here lays the foundation for a reverse-ecology framework for addressing key questions concerning the assembly of host-associated communities and for informing clinical efforts to manipulate the microbiome.

2.2. Introduction

The human body is home to numerous microbial species and several complex microbial ecosystems. Advances in sequencing technologies and metagenomics now allow researchers to characterize the composition of microbial communities that inhabit the human body and the variation these communities exhibit in health and in disease [2,3,45]. Specifically, recent studies of the microbiome have found tremendous variation among healthy individuals [2] and demonstrated clear associations between species composition and several host phenotypes including obesity [46,47], inflammatory bowel disease (IBD) [3], and diabetes [48], as well as with external factors such as diet [49]. These studies further demonstrate that as in many other

ecosystems, community composition in the microbiome exhibits distinct patterns that clearly deviate from a random distribution. Specifically, the human microbiome exhibits a significantly high frequency of *checkerboards*: pairs of species that exclude one another from shared habitats [35,50]. These patterns are similar to those seen in macro-ecological communities, suggesting that similar pressures may act upon these communities [36]. Phylogenetic analysis of community composition in the gastrointestinal microbiomes of domesticated animals similarly revealed that deterministic interactions and niche processes, rather than stochastic neutral forces, dominate community assembly [37].

These studies provide valuable insights into potentially important regularities in the structure of host associated communities. Just as important, however, is to reveal the underlying ecological forces that give rise to such regularities. Identifying these forces and the processes at play in structuring human-associated communities is crucial for developing a principled understanding of the mechanisms that maintain microbiome composition and drive disease-associated compositional shifts, and will ultimately inform clinical efforts to manipulate the microbiome.

Yet, revealing the specific underlying forces that govern the structure of ecosystems and that give rise to specific patterns is a challenging task [36]. Fundamentally different processes and distinct assembly rules can produce similar patterns [51]. Specifically, two alternative processes can account for the observed checkerboard pattern. Diamond [40] suggested a *species assortment* model, in which competitive interactions between species lead to mutual exclusion. Alternatively, a checkerboard pattern can be attributed to a *habitat filtering* model, in which excluding species have affinities for non-overlapping niches [52,53]. Compositional

studies alone therefore may not clearly distinguish a species assortment from a habitat filtering model of assembly, and as a result cannot be used to pinpoint the driving forces that structure communities.

One way to elucidate community structuring forces is to supplement compositional studies with prior knowledge or mechanistic models of the interaction between species [53]. For example, if in a given set of communities, species that exclude one another are known to compete for resources, one could argue that these communities are structured by species assortment. Conversely, knowing that species with similar nutritional requirements tend to co-occur suggests that these species are sorted by habitat filtering [52]. Such information is often available in macro-ecological contexts from phenotypic traits or from feeding habits. In contrast, however, most species of the human microbiota have only recently been identified, and lack a detailed biochemical description of their nutritional requirements and metabolic interactions.

In this chapter, I detail my work to develop a systems biology approach which utilizes genome scale metabolic modeling to address this challenge by augmenting co-occurrence data with computational predictions of species interaction. Specifically, I describe how I developed a *reverse-ecology* analytical framework to obtain insights into the ecology of microorganisms and their environments directly from genomic data and reconstructed metabolic models [54–56]. I then describe how I extended this framework to predict competitive and cooperative interactions between microbes, and how by integrating these mechanistic models of species interaction with co-occurrence data I demonstrated that the assembly of the human microbiome is governed by habitat filtering rather than by species assortment.

2.3. Results

2.3.1. A reverse-ecology framework for predicting species interaction

I used genome-scale metabolic network models to predict the interactions between pairs of microbial species. Networks were reconstructed based on metabolic annotation of available whole genome sequences (Methods). Such network-based models are clearly a simplified representation of the underlying metabolic pathways and dynamics, yet they have proved extremely powerful in elucidating various aspects of microbial metabolism [57,58]. Specifically, the reverse-ecology framework [56] has successfully used such models to predict important ecological attributes, including an organism's biochemical environment [54], its interaction with eukaryotic hosts or with other species [59–61], and ecological strategies for coping with co-habiting species [57] (and see also Refs [58,62] for additional applications). Following this approach, I used the seed set detection algorithm [54] to analyze the metabolic network of each species. This graph theory-based algorithm identified the set of compounds an organism exogenously acquires from its environment, representing the organism's *nutritional profile*. Given the predicted nutritional profile of each species, I introduced two pairwise indices of metabolic interaction (Methods). I defined the *metabolic competition index* as the fraction of compounds in a species' nutritional profile that are also included in its partner's nutritional profile (Figure 2.1 A). This provides a proxy for niche overlap and for the potential level of competition one species may experience in the presence of another. I additionally defined the *metabolic complementarity index* as the fraction of compounds in one species' nutritional profile appearing in the metabolic network but not in the nutritional profile of its partner (Figure 2.1 B). While both species utilize these compounds, one acquires them exogenously while the other synthesizes

them from metabolic precursors, suggesting niche complementarity and potential syntrophy (see Methods for additional details).

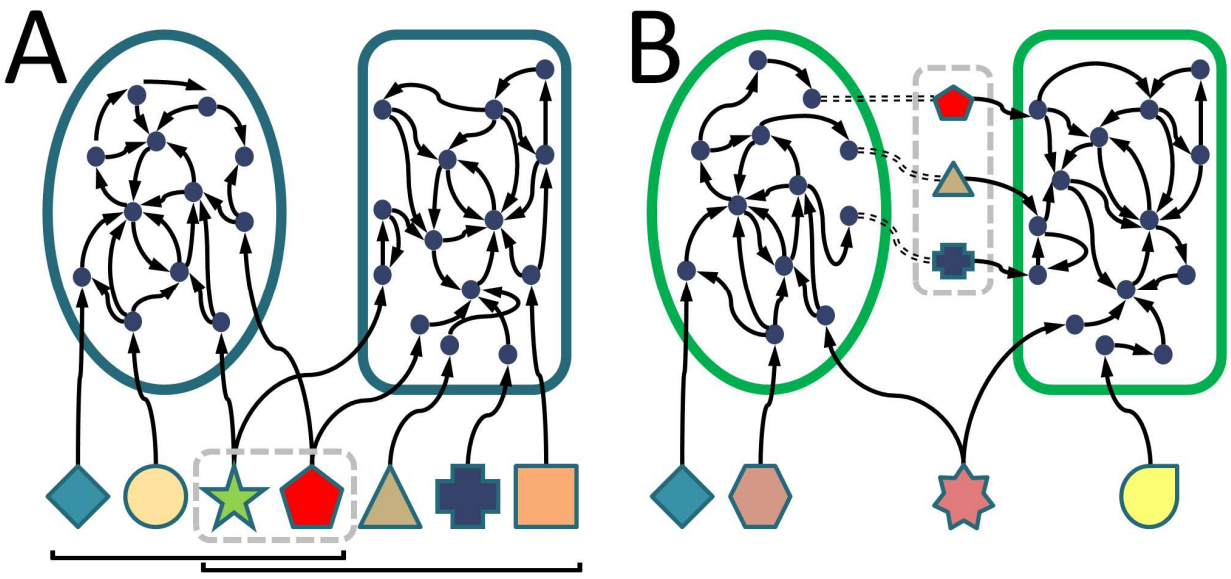


Figure 2.1: An illustration of model-based prediction of species interaction

The metabolic network of each species is reconstructed with nodes representing metabolites and edges connecting substrates to products. The shaped nodes represent exogenously acquired nutrients (seeds). (A) Evaluating metabolic competition. The brackets indicate the 4 and 5 seed nutrients exogenously acquired by the ellipse- and rectangle-shaped species respectively. The 2 metabolites enclosed in a dashed contour denote shared nutrients for which the two species may compete. Accordingly, in this illustration, the competition index experienced by the first species in the presence of the second is 2/4 while the competition index of the second species in the presence of the first is 2/5. (B) Evaluating metabolic complementarity. The compounds enclosed in a dashed contour denote nutrients required by the second species that can be synthesized by the first species. In this example, the complementarity index of the second species in the presence of the first is 3/5.

Contrasting these predicted interaction indices with species co-occurrence patterns allows one to distinguish communities assembled by species assortment from communities assembled by habitat filtering. Specifically, as described above, a negative correlation between co-occurrence and metabolic competition (or a positive correlation between co-occurrence and metabolic complementarity) suggests that community assembly is strongly impacted by species interactions: species that compete for limited resources exclude one another from shared habitats whereas species with complementary (and potentially cooperative) nutritional

requirements tend to co-occur. In contrast, a positive correlation between co-occurrences and metabolic competition suggests community assembly by habitat filtering: a specific environment which offers some set of resources will be inhabited by species that require these resources (and that accordingly have similar nutritional requirements), whereas a different environment (e.g., a different sample) offering a different set of nutrients will select for a different set of species.

2.3.2. Predicted interactions recapitulate species interaction between oral microorganisms

To validate this framework, I employed it to predict metabolic interactions among several species of the human oral microbiome whose interactions have been previously characterized. The human oral microbiome is relatively well described, and many oral species have already been cultured [63]. These species interact via signaling as well as metabolic mechanisms, leading to a characteristic colonization pattern. Late colonizing species are dependent on the presence of early colonizers that attach to the salivary pellicle for survival in the mouth. Pathogens typically arrive later in the cycle, once conditions favorable for their growth are established.

I focused on seven oral species known to influence one another's growth in shared environments (see Methods) (Table 6.1). These species appear during different periods of dental plaque formation, ranging from initial colonizers to late-arriving pathogens [64]. I reconstructed the metabolic networks of these species and determined their nutritional profiles, which I then used to calculate each pair's metabolic competition and metabolic complementarity indices (Methods). Predicted metabolic interaction indices (Table 6.2 and Table 6.3) captured

species' roles within the community and their behavior with interacting partners. Specifically, the pair *S. oralis* and *S. gordonii* have the lowest metabolic complementarity and the highest metabolic competition of all pairs. These two initial colonizers were shown to behave antagonistically [65,66] and are expected to exploit similar niches. Furthermore, in relation to all other species, *P. gingivalis* is the most complemented and poses the least competition to other species, which reflects its ability to grow mutualistically with a wide array of species from all phases of colony formation [66] (and see also 6.1.1).

To further evaluate predicted metabolic interactions on a large scale, I collected from the literature the growth rates of these species alone and in combinations using saliva as a sole nutrient source [63]. To avoid comparison of absolute growth rate across potentially different conditions, I used this data in a comparative manner, generating a list of cases in which a given species was shown to grow better with one species than with another (Methods; Table 6.4). Notably, the well-controlled environments in which these experiments were performed as well as the focus on comparative growth rate analysis allowed me to control for all factors influencing growth of a species (such as habitat heterogeneity) except for the presence or absence of interacting partners. Accordingly, in these growth assays I expected that species would flourish when their interacting partners exploit non-overlapping niches, reducing the detrimental effects of competition. As expected, species that improve growth of the partner tended to have higher metabolic complementarity and lower metabolic competition with those partners ($p < 0.027$ and $p < 4 \times 10^{-4}$, respectively; Methods). A more stringent analysis of this data yielded similar results (see also section 6.1.1). Combined, the findings above demonstrated that metabolic interactions indices successfully reflect the effect of species interaction on growth.

2.3.3. Predicted metabolic interactions and co-occurrences in the gut microbiome

I next turned to investigate species interactions in the gut microbiome. In contrast to the controlled growth assays described above, here, I considered the composition of naturally occurring communities as measured by metagenome sequencing and aimed to elucidate the forces governing the assembly of these communities. Specifically, I focused on a set of 154 prevalent gut species, whose abundances across 124 individuals were obtained from shotgun metagenomic analysis [3] (Methods; Table 6.5). To quantify the co-occurrence of the various species I calculated the abundance-based Jaccard similarity index between all pairs of species (Methods). Using alternative co-occurrence metrics did not qualitatively change the results reported below (see section 6.1.2). I collected genome annotations for all species from IMG [67]. Following the modeling and analysis procedure discussed above, I calculated the metabolic competition and metabolic complementarity indices for all pairs of species (See Methods).

2.3.4. Comparing predicted interactions and co-occurrence patterns suggests that habitat filtering shapes the gut microbiome

I used these data to investigate the association between metabolic interaction and co-occurrence across all samples and all species. Specifically, I wished to determine whether species that compete with one another tend to co-occur or to segregate. I found that the metabolic competition index correlated positively with co-occurrence, whereas the metabolic complementarity index correlated negatively with co-occurrence ($\rho = 0.211$, $p < 10^{-4}$ & $\rho = -0.193$, $p < 10^{-4}$, respectively, Mantel correlation test; see Methods). Notably, while the

correlation is relatively mild, it is extremely significant, with *none* of the permuted null models (see Methods) producing an equal or higher correlation value. This association between metabolic interaction and co-occurrence is even stronger when the analysis is limited to species pairs with coherent interaction indices (see section 6.1.3). As discussed above, these findings suggest that habitat filtering, rather than species assortment, is the dominant structuring force in the intestinal microbiome.

2.3.5. Metabolic interactions of species' partners and excluders

Given this observed correlation, I next sought to determine whether my framework can distinguish species that tend to co-occur significantly with a given species from those that tend to exclude it. For every species in our set, I defined as *partners* those 25% of species with which it has the highest co-occurrence index, and *excluders* as the 25% with which it has the lowest co-occurrence index. Using different threshold values for defining partners and excluders did not qualitatively change the findings reported below (see 6.1.4). I compared the mean competition and complementarity indices of partners and excluders for each species. I found that in 82% of species (127 out of 154; $p < 2x10^{-4}$, permutation analysis; see 6.1.4) the mean competition index with partners is higher than with excluders and that in 86% of species (133 out of 154; $p < 1x10^{-4}$, permutation analysis; see section 6.1.4) the mean complementarity index is lower with partners than with excluders (Figure 2.2). Moreover, separation between partners and excluders is particularly strong when the analysis is limited only to species pairs that exhibit consistent co-occurrence patterns across different host health states (see section 6.1.5). Examining various ecological attributes, I additionally verified that this separation of partners and excluders is consistent across species and does not typify species with any specific ecological label (see section 6.1.6, Table 6.6). I further demonstrated that metabolic versatility

does not explain the observed association between co-occurrence and metabolic competition (see section 6.1.7).

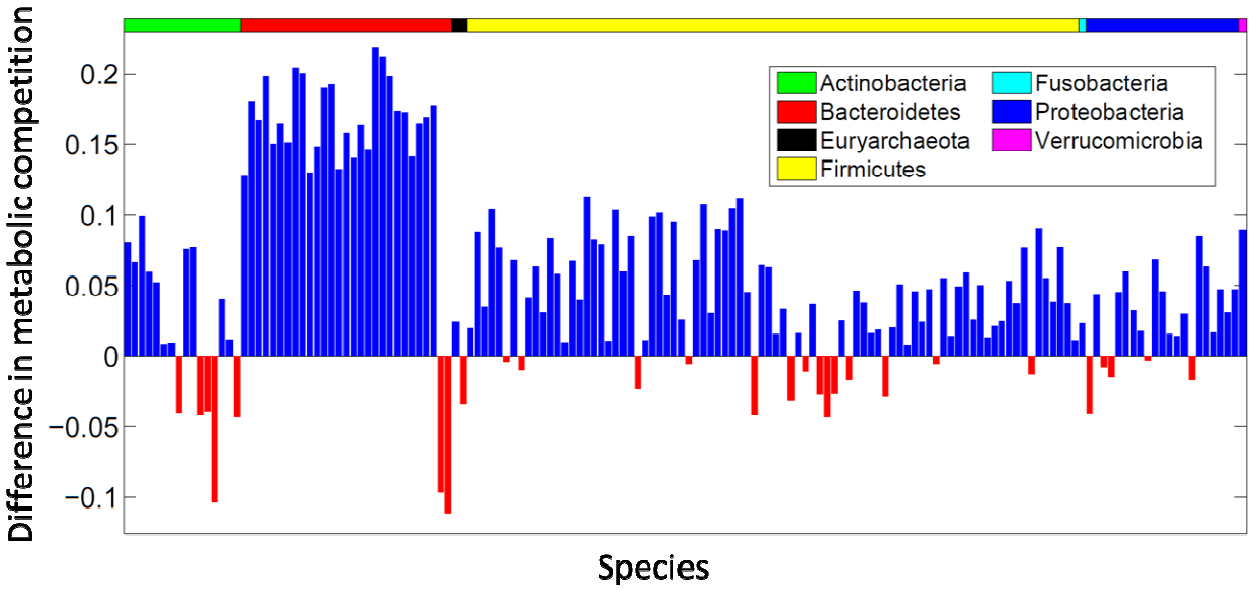


Figure 2.2: Partner species have higher metabolic competition than excluder species

Each bar represents a target species; bar height representing the difference between the mean competition index of its partners and the mean competition index of its excluders. In total, 82% of species have higher metabolic competition index with partners (blue bars).

2.3.6. Habitat filtering in the gut microbiome cannot be explained by the co-occurrence of phylogenetically related species

Previous studies have found that phylogenetically related species tend to co-occur in the gut [3,68]. Similarly, functional capacity and nutritional preferences are strongly associated with phylogeny [54,69]. Subsequently, I next sought to confirm that the above association between co-occurrence and nutritional profile overlap did not derive solely from phylogenetic relatedness. To this end, I used 16s rRNA sequence similarity to quantify the phylogenetic distance between all species in our analysis. I found that metabolic interaction and co-occurrence are still significantly correlated even when controlling for phylogenetic distance (Table 6.7). Thus, while

phylogenetically related species do co-occur in the gut [3], this alone cannot account for the observed habitat filtering signature. To further control for phylogeny, I additionally examined the correlation between metabolic interaction and co-occurrence within each phylum separately. We observed a similar trend, wherein co-occurrence correlates positively with metabolic competition and negatively with metabolic complementarity (Table 6.8). Notably, the magnitude of the correlation between metabolic interaction and co-occurrence within phyla is markedly higher compared to the correlation observed across all species, suggesting that the impact of various structuring forces vary at different phylogenetic scales (see also Discussion).

To further examine the link between metabolic interaction, phylogenetic relatedness, and co-occurrence in detail, I binned all species pairs by both metabolic competition index and phylogenetic distance and calculated the average co-occurrence in each such bin. As demonstrated in Figure 2.3 A, phylogenetic relatedness is correlated with metabolic competition index ($\rho = 0.457$, $p < 10^{-4}$ Mantel correlation test). However, for a given phylogenetic distance, I still observed an increase in co-occurrence as the level of competition increases. To more rigorously validate this finding, I additionally examined whether the competition index with partners (as defined above) differs from the competition index with excluders across different phylogenetic distances. I again found that partners are associated with significantly higher metabolic competition than excluders across all phylogenetic distances (Figure 2.3 B). Additional analysis comparing competition, complementarity, and phylogeny in distinguishing partners vs. excluders can be found in section 6.1.8.

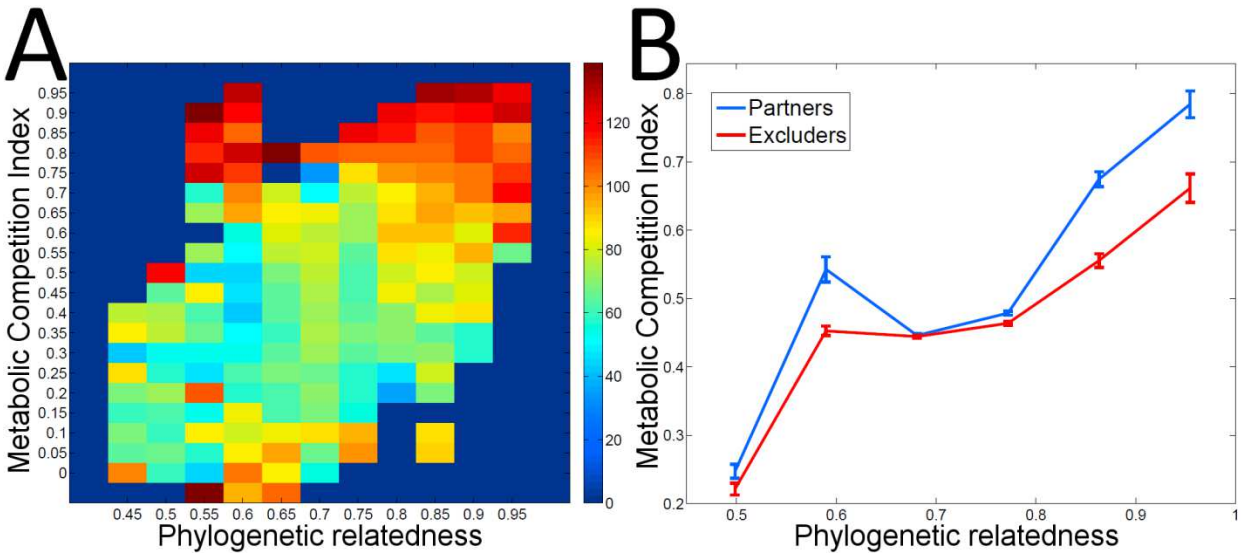


Figure 2.3: Habitat filtering in the gut microbiome across varying phylogenetic distances

Heat map of ranked co-occurrence score, binned by phylogenetic distance (x-axis) and metabolic competition index (y-axis). The color of each bin represents the mean co-occurrence of all pairwise associations within it. Evidently, even amongst species pairs with a given phylogenetic relatedness, mean co-occurrence tends to increase with metabolic competition (see red regions at the top of the heat map). (B) Average metabolic competition index and standard error of partners and excluders v. phylogenetic relatedness. At any level of phylogenetic relatedness, species have more similar nutritional profiles (significantly high metabolic competition index) with partners than with excluders ($p < 0.05$ in all bins, one-tailed Mann-Whitney U test; see also Table 6.9).

2.3.7. Compositional shifts associated with host health and BMI do not fully account for observed habitat filtering patterns

The above findings suggest a habitat filtering model, wherein some properties of the gut environment govern variation in species composition. Notably, previous studies of the gut microbiome identified a strong association between species composition and both obesity [46,47,62] and inflammatory bowel disease [3,62], suggesting these may be major environmental filters influencing community composition. I therefore examined whether these host states can solely account for the observed habitat filtering patterns. To this end, I divided the 124 samples into 4 groups: healthy/lean, healthy/obese, IBD/lean, and IBD/obese. If host state is indeed the sole environmental determinant affecting species filtering, the correlation

reported above between co-occurrence and metabolic interaction should disappear when considering samples from each of these controlled groups separately. I determined the co-occurrence of all species pairs within each group, and calculated again the correlation between metabolic interaction indices and co-occurrence. In all groups, co-occurrence still correlated positively with metabolic competition and negatively with metabolic complementarity (Table 6.10). I similarly found that controlling for additional host attributes, including nationality and enterotype, did not alter this pattern (section 6.1.9). Taken together, these findings imply that the host factors examined do not explain the impact of the host gut environment on the composition of the microbiota and that other (and potentially yet unknown) factors contribute to habitat filtering in the gut environment and to observed species co-occurrence patterns.

2.3.8. Analysis of data from the Human Microbiome Project validates a habitat filtering model

Finally, to validate and extend these results, I examined whether the various patterns reported above can be observed in an additional and independent dataset describing the composition of the human microbiome. To this end, I used data from the Human Microbiome Project (HMP), a large-scale effort to characterize human-associated microbial communities across five major body areas and ~300 healthy individuals [2]. I collected the relative abundances of 335 species (Table 6.11) across 690 shotgun metagenomic samples (Methods). From these data, the co-occurrence of all species pairs was determined. The metabolic competition and complementarity indices of all species pairs were determined as described above.

I first examined the association between metabolic interaction indices and co-occurrence across all samples and all species. As observed above for the intestinal microbiome, co-occurrence correlates positively with the metabolic competition index and negatively with the metabolic complementarity index (Table 6.12), suggesting that the human microbiome is globally structured by habitat filtering. This observation is somewhat expected, given the gross differences between the five major body sites sampled, the distinct characteristic organisms in each [2], and the tendency of species to co-occur across related specific subsites [68]. The obtained correlations are relatively weak but are highly significant (Mantel correlation test; Methods) and further increase when controlling for phylogeny (Table 6.12).

Considering data from intestinal samples alone, I again observed a similar correlation pattern, validating a habitat-filtering model as the dominant assembly mechanism in the gut in this second independent survey. I further examined whether this model represents a general plan for structuring host-associated microbial communities or whether communities in other anatomical sites are potentially subject to different structuring forces. Partitioning samples according to body site and repeating my analysis I found that in communities inhabiting the airways, skin, and the urogenital tract, co-occurrence similarly correlates positively with metabolic competition and negatively with metabolic complementarity (Table 6.12). These correlations remain significant when controlling for phylogeny. In the oral community, the observed correlation was generally weaker, probably owing to relatively low number of genomes available and the pooling of several distinct subsites (section 6.1.10).

2.4. Discussion

Much effort has recently been placed on using co-occurrence to predict interactions of microbial species, either globally [26] or within the human microbiome [2,3,68]. These studies provide valuable insights into non-random structure in community composition but may not be sufficient to pinpoint the underlying forces giving rise to this structure. The framework I presented here, which combines species abundance information with mechanistic modeling of species interactions, renders feasible a more principled analysis of these structuring forces. Specifically, I showed that predicted metabolic interactions correlate with co-occurrence patterns and that species with similar nutritional profiles tend to co-occur, suggesting that habitat filtering is the dominant structuring force of the human microbiome. That is, groups of species which feed on the same compounds are directly influenced by the availability of those compounds in the environment and accordingly co-vary in abundance across hosts.

Clearly, community assembly in the gut is a complex process. Habitat filtering and species assortment are not mutually exclusive in structuring communities [52]. For example, primary consumers of polysaccharides may compete over fiber such as cellulose [70], yet they also release oligosaccharides which are consumed by other species [4]. Our analysis identifies habitat filtering as a principal force, but does not imply that direct species interactions do not play a role. The detrimental effects of competition over nutrients may, for example, be mitigated by the sheer abundance of resources, coupled to the naturally high turnover rate in the intestine. Species may still compete over other resources, resulting in lower overall growth [71].

Previous studies of the composition of the microbiome have highlighted phylogeny as a key determinant of co-occurrence patterns [3,68]. However, my analysis demonstrates that

while phylogenetic relatedness is correlated with both co-occurrence and metabolic interaction, phylogeny does not fully account for the observed habitat filtering pattern. In fact the intensity of the habitat filtering signature increases within phyla, indicating that it may be stronger at finer phylogenetic resolutions. These findings interestingly contrast with recent observations of bacterial diversity in the oral cavity, where significant community structure was demonstrated at the level of genera but not of species [35]. These results may further suggest a strong tendency towards convergent genomic evolution in the gut and potential pressure acting on the evolution of intestinal microbes away from functional diversification [69].

Clearly, however, care must be taken in interpreting these results. Scale, for example, is an important factor and must be taken into account. Considering the variation in pH, nutrient content, oxygen content, and other environmental attributes among the various body sites studied by the Human Microbiome Project, a signature of habitat filtering is probably expected when studying whole-body species co-occurrence patterns: different body sites will clearly select for very different sets of organisms. My findings are in line with previous studies demonstrating that body site has the greatest influence in determining species composition, with less variation observed across individuals [2]. Yet, my analysis of each body site and specifically of the gut microbiome indicates that even when most variation in these factors is controlled, organisms are further filtered on a local scale by as yet undetermined environmental factors.

Specifically focusing on the gut microbiome, I demonstrated that several host phenotypes that have been previously associated with community composition, such as obesity, IBD, and host nationality, are not the sole determining axes along which species are filtered. This suggests subtler environmental and ecological determinants are at play, which do not

mirror host condition. A likely candidate is the biochemical content of the gut, to which host diet is the key contributor. Diet has been demonstrated to be a strong predictor of intestinal microbiota composition [49,72], and may accordingly be the primary link between host macro-ecological state and community composition. Specifically, diets which provide a surplus of nutrients preferred by a subset of the community will increase the abundance of those species, in accordance with a habitat filtering process [73].

Clearly, the models used in this study are a simplification of the underlying biology and have several limitations. First, connectivity-based models and topological analysis cannot fully quantify the strength of metabolic interactions. For example, my method weighs each overlapping compound equally in determining metabolic competition, ignoring the potential contribution of each compound to growth or constraints on reaction fluxes. Similarly, my method aims to quantify the set of compounds both species potentially require, but without prior knowledge about nutrient availability it is hard to determine which compound these species will actually compete for. Notably, constraints based approaches can potentially overcome some of these limitations by explicitly modeling the environment and by incorporating constraints on fluxes and nutrient uptake [74]. However, in contrast to the homology-based networks used in this study, such models require detailed biochemical data and a manually curated reconstruction process and are accordingly not yet available for the vast majority of gut species studied here.

Moreover, it is important to note that while nutrient availability is an important factor, metabolic interactions are not the only determinants of partner preference among microbes. Adhesion, co-aggregation, signaling, and antibiotic tolerance are critical to community

assembly. For example, it has recently been shown that microbes form discrete ecological units that cooperate in the production of antibiotics [75]. As molecular methods improve, multiple 'meta-omic' data types (such as metaproteomic and metametabolomic data) are becoming available, providing insights into such complex inter-species processes. Developing advanced analytic and modeling frameworks that integrate these data types is one of the major challenges microbial ecology currently faces [76,77]. Specifically, modeling and predicting the full range of species interactions, and validating predicted interactions via model systems [78] can dramatically improve our understanding of the microbiome in health and in disease.

Notably, elucidating the assembly rules of the microbiome goes beyond gaining a better understanding of basic ecological processes and has profound clinical implications. Specifically, one of the key challenges of human microbiome research is the development of intervention strategies for driving the intestinal microbiome to favorable states and for microbiome-based therapy [58,79]. In this context, the observation that habitat filtering dominates the assembly of the intestinal community suggests that certain species can be targeted with relatively little concern about their interaction with other members of the community. Similarly, high levels of niche overlap among community members may indicate that dietary supplements may not be precise enough to target species individually. An extended framework for analyzing species interactions within clinical settings could play a key role in the development of microbiome-based treatments. For example, identifying the set of compounds for which species compete could inform dietary-based intervention efforts, safe drug development, species isolation, and colonization studies [55,80]. This study and the framework introduced here are an important first step in this direction, highlighting the opportunities and challenges ahead.

2.5. Methods

2.5.1. Species and community data

I obtained a list of seven oral microbial species from Ref [66]. This list comprises species which have been isolated and have had their growth on saliva assayed. A list of prevalent gut microbial species was obtained from Ref [3]. This list comprises 155 bacterial species for which whole genome sequence is available and that had sequence coverage > 1% in a metagenomic sample from at least one of 124 individuals analyzed (Table 6.5). A set of ecological attributes for each species was obtained from the Prokaryotic Genome Project Tables at NCBI (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi) (Table 6.6).

Abundance data in all metagenomic samples was provided courtesy of S. Dusko Ehrlich [3]. Species abundance was calculated as the sum of sequence length from reads unambiguously mapped to a unique region of a species' genome, normalized by the total length of the unique portion of the species' genome sequence. To account for different sequencing depth across samples, genome coverage was normalized to 1Gbp of sequence. Using this shotgun sequencing-based method to estimate the abundance of each genome in the community provides a natural approach to coupling species abundance data with the genomic data used to reconstruct the species' metabolic networks (see below). For each metagenomic sample, nationality, Body Mass Index (BMI), and health state (IBD / Healthy) of each contributing individual was recorded. For Danish individuals, the enterotype was also recorded. Species abundances were normalized to reflect relative abundances. Species co-occurrence was defined as the similarity in abundance profiles as measured by the continuous Jaccard

similarity index (and see section 6.1.2). We further demonstrated that our co-occurrence measures are robust to the number of individuals sampled (section 6.1.2).

2.5.2. Metabolic network reconstruction

I obtained genomic data for all organisms from the Department of Energy Joint Genome Initiative's Integrated Microbial Genomes project (IMG, http://img.jgi.doe.gov) [67]. For each species, the list of genes mapped to the Kyoto Encyclopedia of Genes and Genomes (KEGG) [81] orthologous groups (KOs) was downloaded. I used these data to reconstruct the genome-scale metabolic network of each species. Networks were represented as directed graphs with nodes representing compounds and edges representing reactions linking substrates to products. A detailed description of the reconstruction procedure can be found in Ref [54].

2.5.3. Analysis of growth data of oral species

Growth rate of species was obtained from several previous studies [63] (and see section 6.1.1), which describe growth assays of multiple oral species in various combinations. I generated a list of all species *trios* for which we can comparatively determine partners' influence on growth (Table 6.4). Specifically, each trio is defined as a target species (e.g., *P. gingivalis*) and two partner species: a *favored partner* (e.g., *A. actinomycetemcomitans*) and a *disfavored partner* (e.g., *F. nucleatum*), such that the target species grows better with the favored partner than with the disfavored partner. I used the paired Student's t-test to confirm that the metabolic interaction indices associated with favored partners are significantly different from those associated with disfavored partners. To validate these results with increased stringency, I additionally used a manually curated dataset, obtaining qualitatively similar results (section 6.1.1).

2.5.4. Predicting metabolic competition and complementarity

I used the seed set of each species as a proxy for its nutritional profile. The seed set represents the minimal set of compounds an organism exogenously acquires to synthesize all other compounds and can be inferred from the topology of its metabolic network using a previously published method [54]. Given these nutritional profiles, two interaction indices were calculated for each pair of species: the *metabolic competition index*, and the *metabolic complementarity index*. The metabolic competition index represents the similarity in two species' nutritional profiles. It is calculated as the fraction of compounds of query species *X*'s seed set that are also present in the seed set of a target *Y*. Since seed compounds are associated with a confidence score (see Ref [54]), this fraction is calculated as a normalized weighted sum. This index provides an upper bound for the amount of competition one species can encounter from another. Using an additional and previously described seed set based metric for competition produced qualitatively similar results (section 6.1.11). The metabolic complementarity index represents the complementarity in two species' nutritional profiles and provides an upper limit for potential syntrophy. To this end, I modified the host-parasite biosynthetic support score [59] to reflect potential complementarity between pairs of microbial species. Specifically, the score is calculated as the fraction of seed compounds of a query species *X* that are producible by the metabolic network of a target *Y* but are not a part of *Y*'s seed set. These may also represent compounds essential to one organism that its partner may provide. Notably, neither of these indices is necessarily symmetric.

2.5.5. Estimation of phylogenetic relatedness

I used the level of sequence similarity between 16s rRNA genes as a proxy for the evolutionary distance between species. 16s rRNA gene sequences for 143 species were

collected from IMG [67] or from the GreenGenes database [82]. For 16s analysis, I followed the procedure described in Ref [69].

2.5.6. Evaluating the correlation between co-occurrence scores and metabolic interaction indices

To calculate the correlation between co-occurrence and metabolic interaction, I generated two matrices: the first lists the co-occurrence scores between all species pairs, and the second lists the predicted interaction index (either competition or complementarity). Since co-occurrence scores are generally symmetric while interaction indices are not (see above), I also generated a symmetric version by replacing each element in the interaction matrix with the mean of each value and that opposite the diagonal. The Spearman correlation between the upper triangles of the co-occurrence matrix and the interaction matrix was calculated. To determine the significance of this association, I employed a permutation-based Mantel test. The rows and columns of the co-occurrence matrix were randomly permuted, preserving species identities (i.e. row and column orders are permuted similarly). For each of 10,000 permuted matrices, I again calculated the Spearman correlation, and the p-value is the fraction of permuted matrices with correlations as great as or greater than the original. To control for phylogenetic relatedness, an additional matrix that describes the phylogenetic relatedness between all species was generated (see above), and the Spearman *partial* correlation of the interaction and co-occurrence matrices, controlling for phylogenetic relatedness, was calculated. Significance was determined using the same permutation approach described above.

2.5.7. Analysis of HMP community data

I obtained Shotgun metagenomic community profiling data from the Human Microbiome Project Data Analysis and Coordination Center (DACC) (http://hmpdacc.org/HMSMCP/). These data represent relative abundance of bacteria and archaea at different taxonomic levels, as determined by MetaPhlAn [83]. MetaPhlAn enables estimation of species abundances and comparison across metagenomic samples of different sequencing depths. In total, 397 species level taxa were classified among 690 samples. Each sample represents one of five major body sites. Since MetaPhlAn does not identify taxa at the strain level, representative genomes were selected from IMG. Where possible, genomes marked 'Human Microbiome Project (HMP) Reference Genomes' were selected. In cases where multiple genomes were available, the genome with the greatest number of KO annotations, and then the greatest number of genes was selected. A list of the 335 species from the MetaPhlAn profile and representative genomes is available in Table 6.11. Abundance of these species in each sample was renormalized in the same method as with the MetaHIT data. Percent similarity of the 16S rRNA gene was used to estimate phylogenetic relatedness as before, analyzing 314 species with representative 16S sequences.

3. Large-scale analysis of functional complementarity in co-occurring microbial genomes reveals global ecological strategies

3.1. Summary

Modern metagenomic sequencing technologies allow for the profiling of microbial communities in diverse environments across the globe. Recent years have seen a surge in efforts to characterize microbial communities, with a particular focus towards microbe-microbe and microbe-environment interactions. In this chapter, I describe work I performed to survey on a large scale a metagenome-derived ecological interaction I termed *functional complementarity*, and to study its relation to community assembly. I define as significantly complementary those pairs of functions which are encoded in distinct microbial genomes, yet which are found together across environments more frequently than expected by chance. I show that functional complementarity is rare globally, but the functions which are most often found to be complemented are those that act at the microbial cell-environment interface, as well as between alternative pathways for energy metabolism. Finally, I show complementary metabolic functions tend to be located at the periphery of metabolic networks, yet are separated by few enzymatic steps. These results indicate that functional complementarity is more common between related functions, and serves to allow microbes to partition niches in order to occupy the same habitat. The functional complementarity analysis I present here represents another analytical framework

that integrates metagenome and whole genome information to address open questions in microbial ecology and community assembly.

3.2. Introduction

Historically, microbes have been studied in isolation within controlled settings, yet naturally occurring microorganisms exist in diverse habitats, form complex communities, and come in direct contact with a diversity of species and environmental factors. Recent advances in metagenomic techniques allow researchers to begin to address questions concerning the ecology of microorganisms previously considered only in the macroscopic realm. Null model and phylogenetic analyses have shown that microbial communities are structured by similar forces as their macroecological counterparts [36,39,84,85]. For example, recent studies have shown that while tree-hole communities are assembled by neutral processes [86], assembly of vertebrate intestinal microbiomes is dominated by niche processes [37].

In particular, much effort has been placed into determining what strategies microbes adopt in order to cope with neighboring species occupying their preferred habitats. Recent experiments using synthetic cooperative communities [87,88] support a hypothesis of distributed metabolism: co-occurring microbes adopt a cooperative strategy wherein one organism metabolizes intermediate products cross-fed by an interacting partner [42,89]. Alternatively, *in-vitro* and *in-silico* co-culture experiments of naturally co-occurring species support the hypothesis of niche partitioning: microbes co-exist by focusing on utilizing a non-overlapping set of environmental resources [17,90]. While both strategies are employed by microbial consortia, the relative contribution of each to the organization of the global microbiome has yet to be determined.

Recent years have seen two analytical frameworks in particular becoming increasingly applied to the study of microbial ecology. The first, comparative analysis of whole microbial genomes, has shown that the ecological similarity of organisms, rather than physical proximity, promotes horizontal gene transfer (or possibly the retention of transferred genes) [91], that genomes are organized into ecologically relevant clusters [43], and that both habitat and phylogeny determine genome function similarity [69]. Nonetheless, because analysis of genome sequence information does not directly take into account the environmental context of the organism, it often cannot distinguish between alternative hypotheses to explain observed structure. The second analytical framework, complex genome-scale modeling, has been developed to offer dynamic predictions of organism behavior in response to environmental or genetic perturbations [92,93], and has further been extended to predict responses to the introduction of interacting species [74,94,95]. Critically, because these models require organism-specific manual curation, their applicability is typically limited to exceedingly small consortia and simple chemical environments.

More recently, analytical frameworks have been introduced which aim to combine the broad applicability of comparative genome analysis with the granularity of complex species-specific models. The previous chapter described how such an approach distinguished between alternative models of community assembly within the human microbiome (See also [96,97]). Here, I present another technique that combines metagenome derived co-occurrence information with whole genome derived functional information, as well as its application to global microbial biogeography. Specifically, I profile *functional complementary* among co-occurring microbes. Complementary functions are those often found together across the microbiome, but which are brought together by a combination of distinct organisms. Unlike comparative

analyses, by considering genomes as distinct units of organization, such an approach underscores specific microbe-microbe interactions, yet can be applied on a large scale in order to profile ecological processes acting globally. I use this framework to investigate the role of genomic complementarity on community assembly by utilizing a global network of co-occurring microbes determined from targeted 16S environmental metagenomic sequences [26,82]. I supplement these data with whole genome sequence information in order to ascertain which pairs of functions are complemented more often than expected by chance. I show that while functional complementarity is rare in nature, most significantly complementary pairs of functions act at the microbe-environment interface (*i.e.*, nutrient uptake and environmental sensing) or between alternative forms of energy production. Finally, I show that complementary functions are functionally related despite a tendency to be found toward the metabolic periphery, indicating such complementarity underlies a strategy of niche partitioning among co-occurring microbes.

3.3. Results

3.3.1. Identifying complementary functions of co-occurring microbes on a global scale

In order to identify pairs of complementary microbial functions on a global scale, I utilized a recently published dataset of co-occurring microbes [26]. This dataset represents the global co-incidence of microbial operational taxonomic units (OTUs) (taxa at the 97% identity threshold) determined from full-length 16S targeted metagenomic surveys. In order to characterize every OTU's functional repertoire, I used BLAST to associate each OTU with its best match from a database of wholly sequenced genomes (Methods). The mapping was

performed in a one-to-one correspondence, such that each OTU was mapped only to its best representative genome and each genome was mapped only to its best representative OTU. Finally, I defined the functional capabilities of each OTU as the Kyoto Encyclopedia of Genes and Genomes (KEGG) modules encoded within the genome (Methods). I used modules rather than orthologous groups (KOs) to determine functional capacity because modules are defined functionally, while orthologous groups are determined by sequence. In this way, organisms that perform the same function via non-orthologous genes are considered functionally equivalent.

Using this strict criterion to map genomes to OTUs, 790 OTUs were assigned a whole genome sequence. I focused my attention on the subset of the incidence matrix that contained co-incidences among these mapped OTUs. Specifically, I considered only genomes mapped to OTUs which appeared in at least two samples, and only samples which contained at least two mapped OTUs. This resulted in a final incidence matrix of 386 genomes across 1145 samples, with over 17,000 co-incidence events. Each sample contained an average 38.6 OTUs, although OTU density roughly followed a power-law distribution.

I used these co-incidence and functional annotation data to determine the number of times each pair of functions was complemented. I defined a complement of the functions *a* and *b* as a single instance of a genome with *a* but not *b* appearing in the same sample as a genome with *b* but not *a* (Methods and Figure 3.1). Of 77,241 possible pairs of functions, we find that nearly half (42%) cannot be complemented by any pair of OTUs considered here (*i.e.*, either there exists no genome with *a* but not *b*, or no genome with *b* but not *a*; genomes with both *a* and *b* cannot complement either). Conversely, of the remainder (pairs of function which can be complemented in this dataset), the vast majority (94%) of function pairs are complemented by at

least one pair of co-incident OTUs (an appropriate pair of genomes are found together in at least one sample).

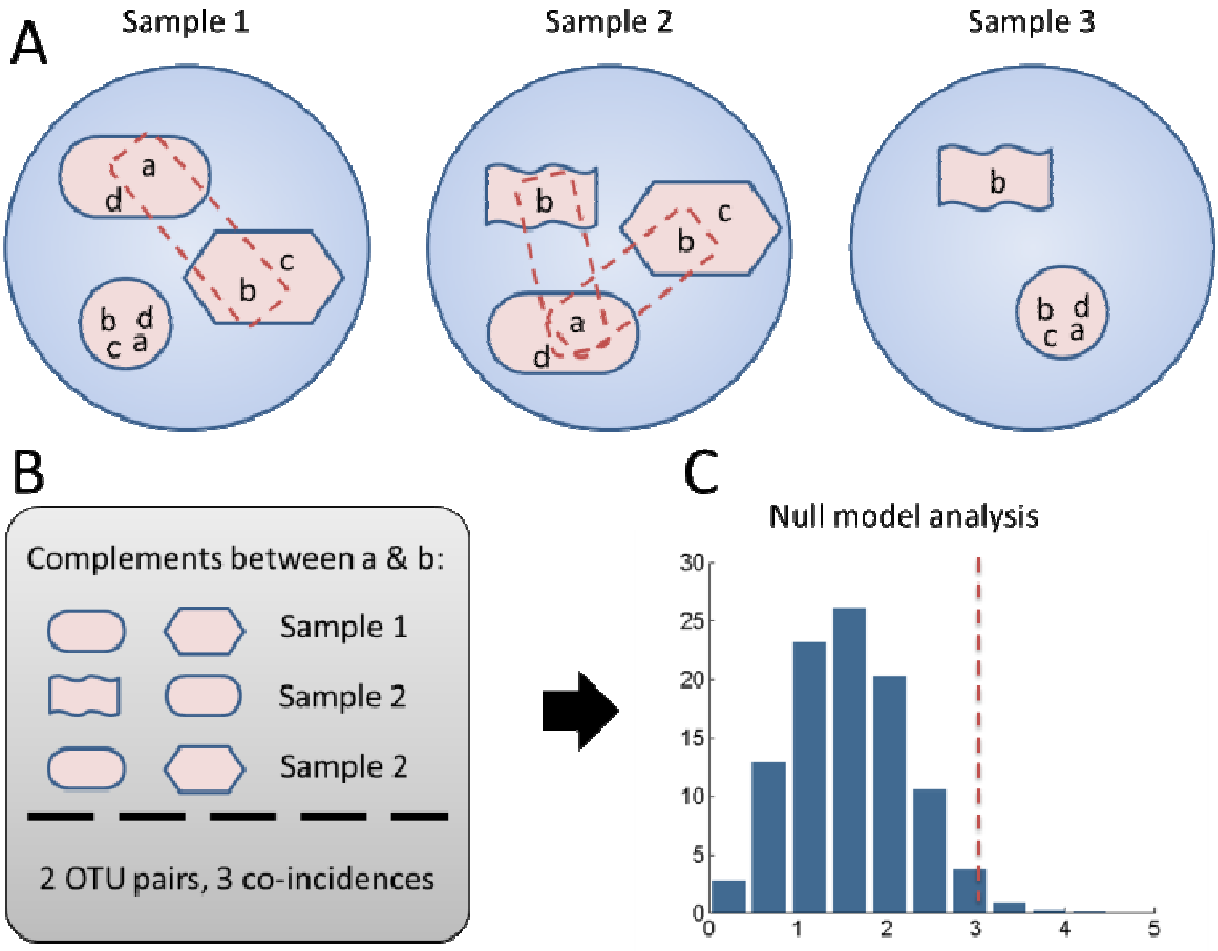


Figure 3.1: An example of the quantification of a single complementarity pair.

In this example, the pair (a, b) is demonstrated. (A) Within each sample, the number of OTUs which contain either function a but not function b or vice-versa are identified. OTUs with both functions or with neither are not considered. (B) The sum of co-incidences between these types of OTUs is the complementarity score (but unique OTU pairs are also counted, see section 7.1.1). (C) The complementarity score is subjected to a null-model analysis. For a more technical description of the approach, including the null-model analysis, see the Methods and Figure 7.1.

Given the observation that most pairs of functions which can be complemented are indeed complemented, I used a null model approach to determine the significance of each

complementarity. I defined a significantly complemented function pair (SCFP) as a pair of functions which are complemented more often than expected by random chance, given the observed taxonomic distribution. To identify these, I permuted the incidence matrix using a method described in Ref. [98] (see Methods). The complete incidence matrix was scanned for 2x2 submatrices forming a 'checkerboard': a pair of OTUs and a pair of samples in which only one OTU is in each sample. Subsequently the incidence of the two OTUs across these samples was switched. Each null matrix was generated by permuting its predecessor 6,000,000 times, until 10,000 matrices were generated. Finally, I counted the number of times each pair of functions was complemented in each null incidence matrix. The significance (*p*-value) of each pair of functions is the fraction of incidence matrices wherein the functions are complemented at least as often as in the experimental case.

Following this approach, I found most function pairs, while complemented at least once, are in fact strongly depleted. Specifically, I found that 52.2% of complemented function pairs have *p*-values greater than 0.9, while 22.2% have *p*-values equal to 1.0 (*i.e.*, the pair is complemented at least as many times in *every* permuted network) (Figure 7.2). The most plausible explanation for this phenomenon is that there exists a distinct set of organisms *A* with function *a* and organisms *B* with function *b,* and typically (but not always) organisms of type *A* segregate from those of type *B*. In order to test this hypothesis, I determined the C-score of the incidence matrix compared to a null distribution [99]; the C-score measures the degree of 'checkerboardedness' of an incidence matrix, a feature characteristic of populations strongly segregated, possibly through the influence of either habitat filtering or competitive exclusion [96,97]. Indeed, I found that the incidence matrix was highly checkerboarded (C-score = 137.10, $p < 4.9995 \times 10^{-4}$, see also Figure 7.3). Given the broad distribution of taxa and environment

types analyzed [26], the most likely assumption is that on a global scale, species are sorted by habitat filtering, which additionally acts to select against functional complementarity. According to this model, pairs of functions which confer an advantage in a given environment will tend to be found in most organisms inhabiting that environment, and almost never in organisms which do not inhabit it. Finally, I observed a number of significantly complemented function pairs. Choosing the significance threshold of $p < 10^{-4}$ (*i.e.*, pairs of functions which are complemented more times in the observed incidence matrix than in *any* null matrix), I found that 1,325 (3.2%) of complemented function pairs are more complementary than expected by chance. This indicates that while most complementarities are strongly selected against, our method was able to identify many significantly complemented function pairs.

3.3.2. Characterizing a network of functional complementarity

I generated a network from these 1,325 SCFPs in which nodes represented functions and edges connected pairs of functions which are significantly complemented. In total, there were 205 functions in this network (Figure 7.4). Analysis of network topology revealed interesting structural properties. Node degree distribution followed a power law, as did the topological coefficient (a measure of the degree to which nodes share neighbors) (Figure 7.5), indicating that while the network contains hubs, these do not cluster together more than other nodes, a characteristic of modular networks [100]. Yet, when compared to a null distribution of permuted networks (Methods), the network had a high characteristic path length (the expected distance between connected nodes; $L_{cp} = 2.514$, $p < 10^{-4}$) as well as a low clustering coefficient (the density of triangle formed between triplets of nodes; $C_c = 0.230$, $p < 0.0125$), indicating that only rarely do we find sets of three functions all of which are complementary to one-another, as expected by transitivity. Additionally, network motif analysis [101] (Figure 7.6) revealed an

enrichment of the 'bi-fan' motif (*i.e.*, a four-node subnetwork where two nodes that do not connect one another both connect to two other functions; $p < 9.99 \times 10^{-4}$) and a depletion of the 'four node chain' (*i.e.*, a four-node subnetwork forming a linear path from the first to the fourth node; $p < 9.99 \times 10^{-4}$). Taken together, these results indicate that this network decomposes into groups of functions which are not directly connected, but which tend to be connected to the same set of functions (see also the discussion of coherence below).

3.3.3. Functional categorization of the complementarity network

I assigned to each function a single broad category term using KEGG pathways and BRITE terms (Methods). Surprisingly, I found a majority of the functions in our network (113, 55%) map to the categories ‘transporter,’ ‘two-component system,’ and ‘phosphotransferase system (PTS),’ processes which act at the microbe-environment interface (*i.e.* functions that are involved in nutrient uptake and/or environmental signal sensing). Beyond that, I found a smaller number of functions are in categories that represent specific biosynthetic pathways, but degradation pathways are all but absent. Only four functions mapped to categories representing degradation pathways (benzoate, dermatan, chondroitin, and keratan degradation), and these connect to only 5 other functions (see also Discussion). Taken together, these results imply a tendency for SCFPs to involve microbe-environment interactions, with a lesser tendency towards biosynthesis pathways.

Notably (and in agreement with motif analysis), I found an appreciable degree of coherence among SCFPs connecting functions which are in the same biosynthetic pathway. For example, 3 of 4 functions mapped to the valine, leucine, and isoleucine biosynthesis pathway category, and each was connected to between 18 and 21 other functions. In total, 23 functions

connected to this pathway, 17 of which (74%) connect to all three functions contained within it. Interestingly, of these 17, the majority were metabolic functions: only 2 were transporters and 2 two-component systems, with no phosphotransferase systems. I also observed a similar degree of coherence within the lysine biosynthesis pathway: of the 8 functions connected to functions of this category, half connected to all 4, and 75% to at least 3. These results support the hypothesis that these pathways tend to be entirely present or entirely absent within a given genome, causing all constituent functions to have similar complementarity profiles [43].

Finally, to generate a summary network that describes the behavior of significantly complemented function pairs at a broad level, I aggregated functions of the same category into metanodes, and determined which metanodes were connected by a greater number of SCFPs than expected by chance (Methods). Considering only category pairs supported by at least 3 SCFPs produces a summary network of 44 edges among 23 metanodes (Table 3.1 and Figure 3.2). 16 edges (36%) involve the three environment-related categories described above, with transporters in particular accounting for 12 edges (27%) in the summary network. The enrichment for SCFPs between these environment-interfacing categories again implies that functional complementarity may underscore environmental selection acting on co-occurring organisms, and may represent a response to ecological pressures. Notable metanodes representing biosynthetic pathway categories included valine, leucine and isoleucine biosynthesis, the TCA cycle, and oxidative phosphorylation. Only three metanodes had self-edges: the TCA cycle, Phosphotransferase systems, and two-component systems.

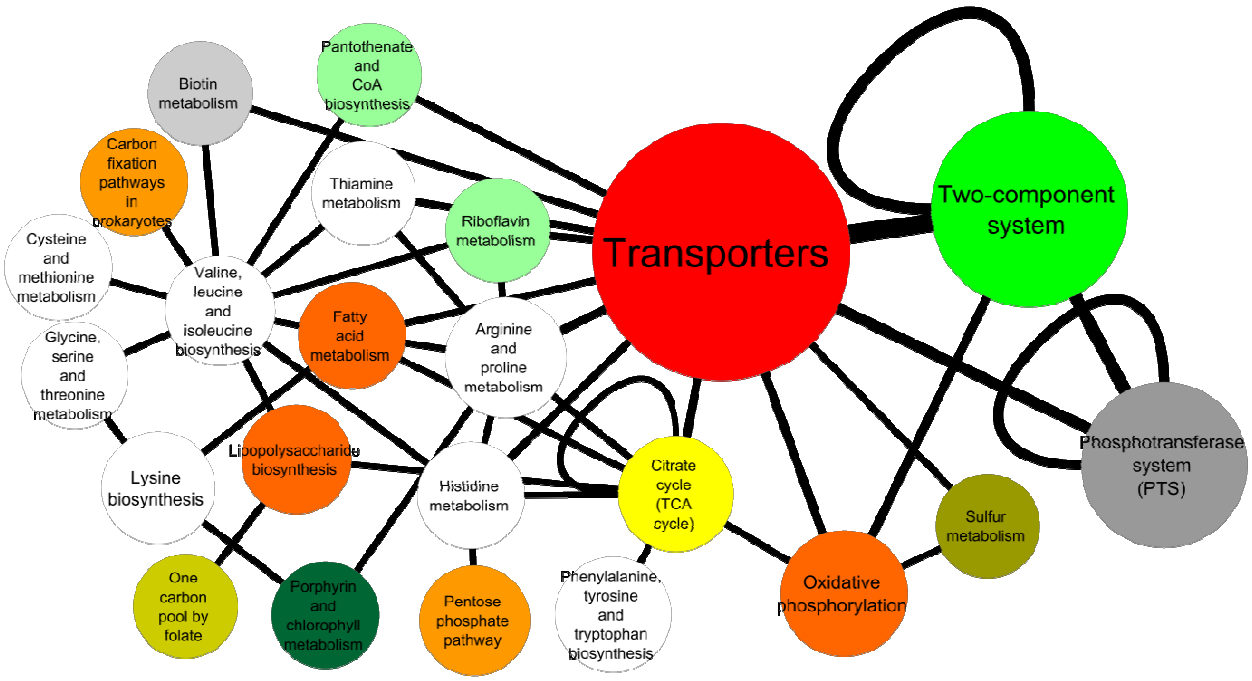


Figure 3.2: Summary network of high-confidence functional complementary

Functions are aggregated according to their category, and node size is scaled according to the number of modules. The largest 3 nodes represent transporters, two-component systems, and phosphotransferase systems (with 56, 34, and 23 functions each, respectively). Edges connect categories with a significant number of complements after FDR correction (FDR < 5%). Edge sizes are scaled by the number of complements between (or within) a given category (with the greatest number of edges between transporters and two-component systems, 175). Only interactions supported by at least 3 complementary function pairs are shown. See also Figure 7.4 and Table 3.1.

Table 3.1: Edge types significantly enriched in the network of functional complementarity

Category	Category	N	K	*p* <	FDR
Arginine and proline metabolism	Fatty acid metabolism	4	18	1.00×10^{-5}	8.38E-04
Arginine and proline metabolism	Histidine metabolism	4	18	1.00E-05	8.38E-04
Arginine and proline metabolism	Porphyrin and chlorophyll metabolism	3	18	2.16E-04	1.10E-02
Arginine and proline metabolism	Riboflavin metabolism	3	9	1.00E-05	8.38E-04
Arginine and proline metabolism	Thiamine metabolism	3	9	1.00E-05	8.38E-04
Citrate cycle (TCA cycle)	Arginine and proline	5	45	1.14E-04	6.16E-03

	metabolism				
Citrate cycle (TCA cycle)	Citrate cycle (TCA cycle)	4	10	0.00E+00	4.70E-05
Citrate cycle (TCA cycle)	Fatty acid metabolism	3	10	1.70E-05	1.23E-03
Citrate cycle (TCA cycle)	Histidine metabolism	3	10	1.70E-05	1.23E-03
Citrate cycle (TCA cycle)	Lipopolysaccharide biosynthesis	3	20	3.32E-04	1.52E-02
Citrate cycle (TCA cycle)	Oxidative phosphorylation	6	60	7.40E-05	4.36E-03
Citrate cycle (TCA cycle)	Phenylalanine, tyrosine and tryptophan biosynthesis	3	25	8.11E-04	3.54E-02
Fatty acid metabolism	Lysine biosynthesis	4	14	3.00E-06	2.71E-04
Histidine metabolism	Pentose phosphate pathway	4	12	1.00E-06	1.25E-04
Lipopolysaccharide biosynthesis	One carbon pool by folate	3	12	3.80E-05	2.48E-03
Lysine biosynthesis	Glycine, serine and threonine metabolism	4	21	2.40E-05	1.66E-03
Lysine biosynthesis	Porphyrin and chlorophyll metabolism	3	14	7.50E-05	4.36E-03
Oxidative phosphorylation	Sulfur metabolism	3	12	3.80E-05	2.48E-03
Phosphotransferase system (PTS)	Phosphotransferase system (PTS)	51	276	0.00E+00	0.00E+00
Transporters	Arginine and proline metabolism	30	810	3.70E-05	2.48E-03
Transporters	Biotin metabolism	9	90	3.00E-06	3.52E-04
Transporters	Citrate cycle (TCA cycle)	26	450	0.00E+00	4.00E-06
Transporters	Fatty acid metabolism	10	180	3.02E-04	1.41E-02
Transporters	Histidine metabolism	14	180	1.00E-06	7.80E-05
Transporters	Oxidative phosphorylation	32	1080	1.22E-03	4.71E-02
Transporters	Pantothenate and CoA biosynthesis	12	180	1.60E-05	1.23E-03
Transporters	Phosphotransferase system (PTS)	68	2160	1.00E-06	1.06E-04
Transporters	Riboflavin metabolism	14	90	0.00E+00	0.00E+00
Transporters	Sulfur metabolism	6	90	9.22E-04	3.72E-02
Transporters	Thiamine metabolism	13	90	0.00E+00	0.00E+00
Transporters	Two-component system	175	6660	0.00E+00	1.00E-06
Two-component system	Oxidative phosphorylation	34	888	7.00E-06	6.23E-04
Two-component system	Phosphotransferase system (PTS)	89	1776	0.00E+00	0.00E+00
Two-component system	Two-component system	72	2701	1.09E-04	5.97E-03

Valine, leucine and isoleucine biosynthesis	Biotin metabolism	3	3	0.00E+00	0.00E+00
Valine, leucine and isoleucine biosynthesis	Carbon fixation pathways in prokaryotes	3	15	1.00E-04	5.62E-03
Valine, leucine and isoleucine biosynthesis	Cysteine and methionine metabolism	3	18	2.16E-04	1.10E-02
Valine, leucine and isoleucine biosynthesis	Fatty acid metabolism	3	6	1.00E-06	1.37E-04
Valine, leucine and isoleucine biosynthesis	Glycine, serine and threonine metabolism	3	9	1.00E-05	8.38E-04
Valine, leucine and isoleucine biosynthesis	Histidine metabolism	6	6	0.00E+00	0.00E+00
Valine, leucine and isoleucine biosynthesis	Lipopolysaccharide biosynthesis	6	12	0.00E+00	0.00E+00
Valine, leucine and isoleucine biosynthesis	Pantothenate and CoA biosynthesis	3	6	1.00E-06	1.37E-04
Valine, leucine and isoleucine biosynthesis	Riboflavin metabolism	3	3	0.00E+00	0.00E+00
Valine, leucine and isoleucine biosynthesis	Thiamine metabolism	3	3	0.00E+00	0.00E+00

3.3.4. Complementarity demonstrates alternative lifestyles but is not distributed across whole pathways

I more closely examined significant complements between functions of the same category. These SCFPs might indicate distributed metabolism among microbes, where a larger pathway is reconstituted from subcomponents in separate organisms. We find only 5 categories which contain SCFPs within themselves: transporters, two-component systems, PTSs, oxidative phosphorylation, and the TCA cycle. Notably, only the last two are biosynthesis pathways, representing an extreme minority (7 out of 217 intra-category SCFPs, 3.2%). This demonstrates that while co-occurring bacteria may exhibit distributed metabolism [102], functional complementarity most commonly represents microbes interacting in distinct ways with their shared environment.

In order to investigate whether these metabolic intra-category SCFPs were the result of distributed metabolism, I more closely investigate these function pairs. Specifically, I determined which functions were complemented, which taxa provided each function, and in which environments these taxa were found together. Within the TCA cycle, I found two major complementarities in effect (Figure 7.7). The first is between a single pyruvate oxidation reaction and three functions comprising the Krebs cycle; notably, many of the organisms contributing pyruvate oxidation to this complement are of a clade known to lack the TCA cycle [103]. The second is between the first and second carbon oxidation steps in the TCA cycle, coinciding with alternative approaches in applying the TCA cycle in a purely biosynthetic manner. Notably, many complements between these functions occurred in the mammalian gut, a hypoxic environment in which many organisms ferment carbohydrates anaerobically. Similarly, within the oxidative phosphorylation pathway I found 3 complements, each between cytochrome aa3-600 (menaquinol oxidase) and a component of the electron transport chain (8). Notably, all contributors of cytochrome aa3-600 are of the lineage bacillales, which are known to maintain an atypical electron transfer chain [104–106]. Thus, functional complementarity within core metabolism more commonly arises through the co-incidence of microbes with distinct metabolic programs rather than through distribution of metabolic functions.

3.3.5. Complemented functions are peripherally located and functionally related

Given this lack of a signature of distributed metabolism, I investigated the extent to which functional complementarity contributes to niche partitioning. I used a network model of global microbial metabolism in order to quantitatively characterize the functional relatedness of complementary functions. In this network, nodes represent compounds and a directed edge

connects substrates to products. The metabolites associated with each KEGG module were determined (See Methods). Finally, I determined the average betweenness centrality (the fraction of shortest paths between connected nodes which pass through a given node) of all compounds associated with each pair of functions. I found that compounds involved in significantly complemented function pairs have lower mean centrality than other function pairs ($p < 9.01 \times 10^{-9}$, one-sided Wilcoxon rank-sum test, Figure 7.9 A).

Given the tendency of complementary functions to exist on the periphery of the metabolic network and the relative rarity of complementary function pairs within a single pathway, I next investigated the functional distance between pairs of complemented functions (Methods). A pair of peripheral functions is potentially separated by greater distance if the functions are found on opposite sides of the network, as would be expected of *unrelated* peripheral functions. Surprisingly, I found that pairs of significantly complemented function pairs are in general located closer to one another within the network ($p < 10^{-300}$, one-sided Wilcoxon rank-sum test, Figure 7.9 B). Similarly, I tested the number of enzymatic steps taken before pairs of SCFPs converge on a common product (Methods); I found that significantly complementary functions converge in fewer steps than non-complementary functions ($p < 0.0395$, one-sided Wilcoxon rank-sum test, Figure 7.9 C).

Metabolic network proximity implies functional relatedness, but doesn't guarantee that complementary functions act on chemically similar substrates. In order to directly quantify chemical similarity I determined the maximum common subgraph of all pairs of compounds in the metabolic network [107]. For any given pair of molecules this measures not only the number of atoms they have in common, but the largest number of atomic bonds in the same

configuration. From this, the Tanimoto coefficient of similarity [108] for the chemical compounds was calculated. I found that significantly complementary functions act on chemically similar compounds ($p < 10^{-300}$, one-sided Wilcoxon rank-sum test, Figure 7.10). Thus, complementary functions tend to be found at the network periphery, and tend to act on related substrates to produce common metabolic products. Taken together, these results support a niche partitioning model of community assembly: microbes inhabiting the same environment exploit a non-overlapping set of resources utilized by common intracellular processes.

3.3.6. Functional complementarity reveals niche partitioning at the local scale

The scale at which a metacommunity is defined greatly impacts the appearance of ecological structure. Because I analyzed global microbial biogeography across a wide array of environmental contexts, the metacommunity was defined by the broadest possible geographic and environmental scales. At this scale, species and functions may appear to cluster together more so than when considered against a narrower environmental context [109]. In order to investigate the impact of scale on the above reported results, I used Environmental Ontology (EnvO) terms investigate niche partitioning within a set of narrower environmental contexts. Specifically, I defined three dominant metacommunities representing the most common environments in our dataset: soil, aquatic, and gut, (comprising 188, 159, and 63 samples respectively; Methods), within which I repeated the previous null model analyses.

Compared to the global metacommunity, at any significance threshold, fewer significantly complemented function pairs were found in all environmentally defined metacommunities. This may be due to reduced taxonomic and functional richness in this subset

of samples, or it may be an indication that at this scale, communities are comparatively less clustered. Functional richness, however, was not greatly reduced within these samples, while average β-diversity was (all $p < 10^{-9}$, Wilcoxon rank-sum test; Table 7.2). This suggests that metacommunity scale, and not functional richness, is the reason fewer functional complements are significant. Subsequently, for these metacommunities I relaxed the significance criteria for SCFPs to the more permissive *p*-value of 0.05. At this threshold all 3 metacommunities contained a markedly distinct set of SCFPs that were distinct from the set of SCFPs found at the global scale (Figure 3.3).

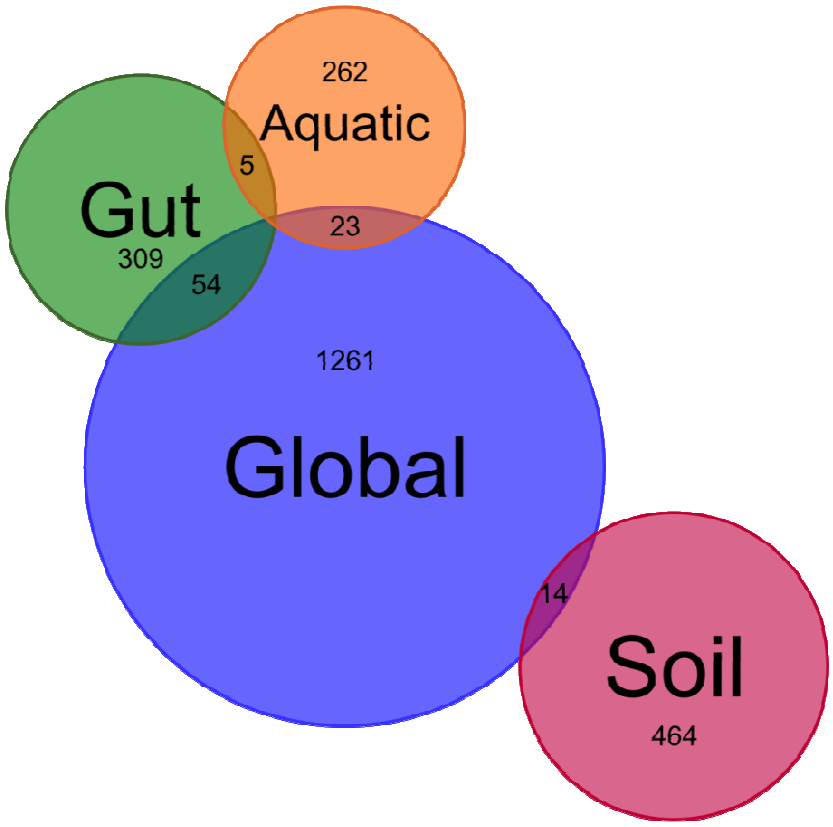


Figure 3.3: Significantly complemented function pairs differ greatly depending on metacommunity scale

The Euler diagram shows the number of SCFPs for each metacommunity, as well as how many are shared between metacommunities. 1 SCFP shared among the global, gut, and aquatic metacommunities is unlabeled for clarity.

Nonetheless, many characteristics of niche partitioning were observed at this scale. Most critically, the substrates of SCFPs in these environments were more chemically similar than other pairs' ($p < 10^{-300}$, Wilcoxon rank-sum test), although centrality, distance, and convergence time differed significantly between SCFPs and null in only some cases (Table 7.3). As I had done before for the global metacommunity, I performed enrichment analysis of SCFPs between and within categories. In the gut and aquatic metacommunities, all but 1 (of 31) such enrichments involved the microbe-environment interfacing categories discussed above (*i.e.*, transporters, two-component systems, and PTS systems), and only these categories were enriched for SCFPs in the same category. The network of complements significant in the soil metacommunity, in contrast to the others, was dominated specifically by two-component systems (which in this case was the only category with a self-enrichment). Most of these SCFPs decomposed into coherent groups of systems involved in nitrogen fixation, photosynthesis, and host-cell invasion (Figure 7.11), non-overlapping niches within the rhizosphere occupied in turn by soil resident rhizobiales, actinobacteria, and mycobacteria [110–114]. These results indicate that at different scales, niches are defined along different environmental axes, and in some communities may be more clearly identified by environmental signal processing than by metabolic capabilities.

3.4. Discussion

In this chapter, I presented a global survey of functional complementarity among co-occurring microbes. Complementary functions are those which are frequently found together within the environment, but which are brought together by a distinct set of co-occurring microbes. Functional complementarity is most common at the microbe-environment interface, but also occurs between alternative energy metabolism pathways. Complementary functions

tend not to reconstitute larger metabolic processes, but instead distinguish alternative lifestyles of co-occurring microbes. Nonetheless, complementary functions tend to be related, as determined by substrate similarity and proximity within the metabolic network.

Given the structure of microbial taxonomic distribution, functional complementarity appears to be rare. Almost half of all pairs of functions do not segregate across genomes, and those that do also strongly segregate across environments. As both genome structure and taxonomic distributions are the result of adaptive and ecological processes, this likely reflects underlying constraints selecting against functional complementarity. Conceivably, a function beneficial in or adapted to a given environment would become common among occupants of that environment (perhaps encoded by non-homologous genes in different taxa), but those organisms would likely be rare in dissimilar environments. This underscores a strong tendency for environmental selection and habitat filtering, leading to the segregation of complementary genomes on a global scale. Accordingly, it has been shown that broad definitions of metacommunity scale make communities appear phylogenetically clustered [109], commonly seen as evidence of habitat filtering [39]. Consequently, a statistical framework which is robust to scale becomes necessary to distinguish the effects of habitat filtering and niche partitioning in community assembly.

Nonetheless, these results indicate a reliance on niche partitioning in microbial community assembly as the predominant strategy to cope with co-occurring microbes occupying the same habitat. Specifically, it reveals that co-occurring microbes tend to extract from their environment non-overlapping sets of nutrients, yet process these in order to operate similar

downstream metabolic processes. For example, a pair of co-occurring microbes may import distinct sets of polysaccharides, but inevitably ferment these using the same metabolic program.

Interestingly, it has been shown that naturally co-occurring microbes sense the appearance of other taxa and respond by activating non-overlapping metabolic inputs [17]. These two phenomena may represent processes operating at different timescales by which complementarity and niche partitioning may eventually lead to specialization. Metabolic inputs which an organism deactivates on the short term may be lost through reductive evolution in the long term, encoding this sort of functional complementarity directly within the genome. While to my knowledge, no experimental co-evolution study has yet demonstrated the long-term result hypothesized here (loss of a pathway to utilize a particular resource), experiments have shown that over the course of generations, complementarity may in fact be encoded within the genome of co-occurring ecotypes. For example, evolution of an initially isogenic *E. coli* population diverged into two populations distinguished by the rate of diauxic shift from glucose to acetate [115]. Furthermore, whole genome sequencing of evolved strains demonstrated causative mutations and evolutionary dynamics [116]. As discussed before, the definition of *function* in an analysis of functional complementarity is particular to the study. If one were to define function in this case as the metabolic reaction norm employed by a strain [115,117], which is encoded by a number of nucleotide polymorphisms, this would in effect be an instance of functional complementarity emerging through genome evolution.

Notably, degradation pathways were all but absent from the network of significantly complementary function pairs. Specifically, only four functions involved in seven significant complements were found. Three of these functions, (dermatan, chondroitin, and keratan

degradation) all act on glycosaminoglycans. Bacteria of the vertebrate guts are known to degrade glycosaminoglycans such as these as an energy source [118]. Beyond that, they are sensed by pathogens and may be critical to crossing the luminal barrier [119]. Therefore, while in eukarya these may represent metabolic end products, it is likely that in the context of microbes, particularly host-associated bacteria, these too are metabolic inputs, and act at the cell-environment interface like other complemented functions.

Choice of carbon source appears to play a significant role in niche partitioning, exhibited by the prevalence of PTS systems and sugar-specific ABC transporters in our network. Nonetheless, closer investigation of inorganic compound transporters imply that niche partitioning is not exclusive to this subset of the chemical environment. For example, complementarity between Fe^{3+} and peptide/Ni transporters is found across a number of environments, including heavy-metal contaminated sites. Notably, known Proteobacterial iron oxidizers (particularly β-proteobacteria) [120] dominate the iron component, with more diversity across the nickel component (including many lactobacillales with potentially no iron requirement [121]). This demonstrates potential partitioning between chemolithotrophs and chemotrophs requiring nickel, a cofactor of carbon monoxide dehydrogenase. Furthermore, iron complex transporters were also complementary to nitrate transporters, but not in environments including *Thiobacillus denitrificans*, which requires nitrate to oxidize iron (potentially indicating that niche overlap drives these organisms into different habitats).

As with any analysis of biogeography, considerations of scale must be taken into account. Alternative definitions of background metacommunity strongly influenced the significance of functional complements, as communities which appear clustered when

compared to a global metacommunity may not appear so within a specific ecological context. Nonetheless, emergent properties of complementarity networks were largely consistent. In all contexts analyzed, most functional complementarity occurred at the microbe-environment interface, specifically between chemically similar substrates. Interestingly, functional complementarity appeared to separate along different dimensions of niche space at different scales. While at the global scale, carbon source appeared to be the major consideration, within the rhizosphere, interactions with resident flora appears to play a major role in structuring the community. Specifically, nitrogen fixing rhizobia complement photosynthetic actinobacteria. It is interesting to note this functionality is not completely segregated across these clades. Depending on environmental conditions, *Nostoc spp.* adopt alternative nitrogen-fixing or phototrophic cell states, [112]. Thus, an organism's choice of ecological strategy is context dependent, chosen from a larger repertoire of functional capabilities. The ability to partition niches along these axes, therefore, is more strongly demonstrated by signaling pathways in these organisms than by metabolism strictly. Indeed, while photosystem-II was significantly complementary to nitrogen fixation, a greater number of complements were found among the two-component systems themselves.

The vast majority of the microbial tree of life remains uncharacterized. While thousands of organisms have had their genomes sequenced, ~90% of these are from only 4 bacterial phyla [122]. Thus relative to the scope of available data, few unique OTU-genome pairs could be mapped in this study (~1,000 out of ~40,000 identified OTUs). This is a significant limitation not only of the work presented here, but of any attempt to apply omic-based analysis of natural microbial communities at the global scale. For this reason, this study focused on emergent properties of community assembly, which are unlikely to be unique to any subset of branches of

the tree of life. Nonetheless, efforts are underway not only to sequence representatives of more taxa but to characterize their functional capacity at the molecular level, which will simultaneously increase the applicability of approaches such as these as well as allow for more focused study of particular systems or communities [123].

Finally, experimental synthetic ecology has recently been introduced as a new framework to investigate the organization and evolution of microbial communities [42,87], and represents a potentially consilient approach to large scale genome analysis. A recent survey demonstrated high potential for distributed metabolism among engineered E. coli strains and presented the hypothesis that metabolic complementarity is common among microbes and serves to reduce the burden imposed by synthesis of costly amino acids [42]. As these results may appear to contrast with those presented here, further discussion is warranted. Notably, while in this study co-cultured partners survived in conditions unsupportive of independent growth, previous systems demonstrated such mutual obligatory cooperation may be based off interactions more akin to scavenging than syntrophy [87]. Furthermore, while whole-genome analysis revealed patterns of auxotrophy associated with phylogeny, this does not reveal how bacteria cope with auxotrophy outside of laboratory settings; they may only persist in environments where essential metabolites naturally occur (e.g., eukaryotic hosts). Future work aimed at the synthesis of experimental and computational synthetic ecology with phylogenomic and metagenomic analysis stands as the most promising route to completing the circuit from observation to hypothesis and onwards to model building and empirical validation.

3.5. Methods

3.6. Incidence and whole genome data

I obtained a list of OTU-sample incidences from Ref [26]. This represents the appearance of all 298,591 full length 16S sequences in the Greengenes database with complete annotations for the fields “author,” “title,” and “isolation_source.” Each sample was defined as a unique combination of these three annotation fields. 16S sequences were aligned using Infernal [124], and hierarchical clustering was performed using the RDP clustering tool [125]; for our analysis, OTUs were defined at the 97% identity threshold. The assignment of Environmental Ontology (EnvO) terms was also taken from [26]. Assignments were made by matching samples' "isolation_source" to EnvO terms (or *exact* or *narrow* synonyms).

Whole genome sequence data was collected from the Department of Energy Joint Genome Initiative’s Integrated Microbial Genomes project (IMG, http://img.jgi.doe.gov) [67] (version 350). Only genomes with sequencing status “finished” were considered. For each genome, the list of genes mapped to the Kyoto Encyclopedia of Genes and Genomes (KEGG) [81] orthologous groups (KOs) was downloaded, as well as gene sequence for any gene wherein the field “Gene Symbol” matched “16S”.

3.6.1. Mapping OTUs to whole genomes

To assign genomes to OTUs, I followed the procedure described in [26]. For each OTU, the 16S sequence with the lowest sum of squares distance to others in the OTU was chosen as a representative sequence. A BLAST database was generated of the 16S sequences from IMG; only sequences length 700bp or greater were included, and only the longest in a given genome. Representative OTU sequences were BLASTed against this database using parameters

described previously [26]. Only alignments of length at least 800bp and with sequence identity at least 97% were considered. Finally, OTUs were mapped uniquely to genomes: All BLAST hits were ranked by bit-score, and OTUs were assigned to the best matched genome which lacked a higher ranking match to another OTU.

3.6.2. Functional annotation of genomes

For this analysis, I defined a “function” as a KEGG module encoded in a genome (e.g., glucose to pyruvate conversion). In the KEGG framework, modules are defined at finer resolution than whole pathways (e.g., gluconeogenesis), but broader than orthologous groups (KOs) or enzymes (e.g., hexokinase). Correspondingly, different organisms might operate the same module using different KOs. Therefore, each KEGG module can be defined as a number of components, and each component can be expressed by a Boolean expression of constituent KOs. In this way, modules can represent functional equivalency of taxa which perform similar functions using non-orthologous genes.

I used the presence of KOs and these Boolean expressions to determine the fraction of components of each module encoded in each genome. I considered a module present in a genome if the genome contained KOs for at least 60% of the components of that module; it was absent from any genome which contained less than 20% of components. Optional components were disregarded. The choice of threshold did not significantly impact functional annotation of genomes (Figure 7.12). Finally, two functions (Ribosome and RNA polymerase) were ignored; these functions are universal among bacteria and their appearance in functional complements could be attributed to annotation errors.

3.6.3. Complementarity search

To determine the complementarity of a pair of functions, I inspected function presence and absence in all OTU co-incidences. The complementarity score of a pair of functions *a* and *b* was defined as the number of times an OTU with function *a* but without function *b* appeared in a sample with an OTU which lacks function *a* and has *b*. Notably, in this scheme the same OTUs appearing together in many samples could contribute considerably to this score, as would numerous OTUs in even one sample. I additionally implemented an alternative approach wherein each co-incident pair of OTUs was only considered once and found qualitatively similar results (see section 7.1.1, Table 7.1, and Figure 7.13).

3.6.4. Null model analysis

I performed the complementarity search on a reduced incidence matrix of only OTUs mapped to genomes and only samples which contained at least 2 such OTUs. The final result was an incidence matrix of 386 OTU-genomes across 1145 samples. To create a distribution of null matrices, I employed a permutation approach (SIM9) from reference [41,98] (Figure 7.1). The incidence matrix was scanned for checkerboards: a pair of OTUs and a pair of samples in which the first OTU appears in the first sample but not the second and vice versa. The submatrix does not need to be contiguous within the complete incidence matrix. The incidence of the OTUs is permuted such that the first OTU now appears only in the second sample and the second OTU appears only in the first sample. Notably, this approach maintains marginal distributions; species richness of samples and OTU prevalence across samples is preserved. To determine the significance of a pair of functions, the complementarity score of that pair was determined in each null incidence matrix. The *p*-value was calculated as the fraction of

incidence matrices in which the complementarity score was as great as or greater than in the unpermuted matrix.

3.6.5. A network of complementary pairs of functions

To generate a network of significantly complemented function pairs, I performed the null model approach described above to all pairs of functions. 1,325 function pairs were found to be significant at the most stringent threshold ($p < 10^{-4}$; the function is complemented more in the experimental matrix than in *any* null matrix). 1113 (82%) of the function pairs in this network have a complementarity score of 100 or greater, indicating the high degree of support for complementarities in this network.

3.6.6. Defining node functional category

As defined by KEGG, modules can be subcomponents of multiple systems including pathways (e.g., phosphotransferase system) and BRITE terms (e.g., transporters). I unambiguously assigned to each module one category in the following process. First, any module with only one pathway or BRITE term was assigned to that category. For modules with multiple possible categories, priority was given to pathways and KEGG terms prevalent in the annotation. ko02060, “Phosphotransferase system (PTS),” the most prevalent pathway among unambiguously categorized modules, was given priority. Subsequently, the most prevalent BRITE term, ko02000, “Transporters,” was chosen. Remaining modules with “Two-component system” in any linked pathways or BRITE terms were categorized as such. In this manner, modules were placed in the category in which the most modules were already placed; ties were determined by assigning the pathway with the lowest ID (presumed to be of the broadest

scope). This protocol failed to categorize only one module, "indolepyruvate ferredoxin oxidoreductase," which was placed in its own category.

3.6.7. Analysis of the complementarity network

Network topology (*i.e.*, characteristic path length, clustering coefficient, node degree and topological coefficient distribution) was characterized using Network Analyzer [126]. To generate a distribution of null networks, I implemented the permutation algorithm described in [101]. Motif analysis was performed using mFinder [127]. Enrichment of 3- and 4-node motifs was determined considering edges as undirected. 3-node motifs were compared against a distribution of 10,000 null networks, 4-node motifs were compared to 1,000. To generate the summary network, the hypergeometric test was used to determine category pairs complemented by more function pairs than expected by chance (Benjamini-Hochberg False Discovery Rate < 0.05). Only category pairs sharing at least 3 complements were considered.

3.6.8. Calculating metabolic centrality and functional distance

To generate a global metabolic network, I used a method previously described [54]. Specifically, nodes in the network represent metabolites and directed edges represent enzymatic reactions. For each reaction, an edge is drawn from all substrates to all products. The list of enzymatic reactions was taken from those in KEGG mapping to any KO present in either a bacterial species from KEGG or IMG. The metabolite content of 215 pathway modules was taken directly from KEGG. For 39 modules lacking an explicit compound list, metabolites were determined from the non-optional KOs by the same method used to generate the metabolic network. 88 modules with the tag “Substrate:” in the comment were associated with

the listed compounds. Finally, 40 modules with the tag “Signal:” in the comment were associated with the listed compounds.

To calculate centrality and distance of modules, the metabolic network was considered undirected, and only the largest connected component was considered. The betweenness centrality of a given node was calculated as the fraction of shortest paths between all nodes in the network which pass through the node [128]. Because modules may contain many compounds, the centrality of a module was taken as the mean centrality of all compounds associated with it. To calculate the distance between a pair of modules, the mean shortest path from each compound in one module to each compound in the other was used. When calculating functional convergence, edge directionality was taken into account. The convergence of two compounds is the minimum number of steps it takes to reach a common compound from both starting nodes. Nodes that do not converge are ignored in calculating the mean convergence time between a pair of modules.

To calculate chemical similarity of a pair of compounds, the maximum common subgraph of the two compounds was determined using SMSD and the number of atoms shared in the common substructure was counted [107]. From this, the Tanimoto coefficient of similarity for each pair was determined [108].

4. Metagenome assembly rules revealed by genetic co-occurrence

4.1. Summary

The prokaryotic genome is highly structured, as evidenced by strong patterns of genetic co-occurrence across the microbial tree of life. The metagenome of a microbial community is a weighted combination of such genomes, yet it is unclear to what extent the metagenome exhibits the same structure as the genome. In this chapter I present work comparing the co-occurrence structure of microbial genomes and metagenomes. I found that the salient feature of genome structure – co-occurrence among related but not directly interacting genes – is largely maintained at the community level. Nonetheless, differential analysis of genome and metagenome composition revealed that co-occurrence structure is lost among core cellular processes such as genetic information processing. In contrast, environmentally facing functions that act at the microbe-environment interface, which define an organism's niche, are relatively more structured in the metagenome than in the genome. Finally, analysis of functions which are cross-coordinated in metagenomes implicates distribution of xenobiotic degradation processes among distinct community members. These results highlight the contrasting evolutionary and ecological forces defining structure of genomes and metagenomes.

4.2. Introduction

Genome-scale shotgun sequencing has transformed the study of microbiology. The number of fully sequenced microbial genomes available allows researchers to characterize the compositional structure of the prokaryotic genome. Structure is often evidenced by non-random patterns of genetic co-occurrence and is largely the result of vertical co-inheritance [44]. Nonetheless, phylogenomic analysis revealed that the gain and loss of functionally related genes is coordinated, and reflects functional relationships not limited to direct molecular interactions [43,129]. Co-occurrence structure can also reveal constraints on genome evolution during adaptation of prokaryotes to specific environments [54,130].

Sequencing has similarly influenced the study of microbial ecology. Using shotgun metagenome sequencing, near-complete assembly of genomes from naturally occurring, complex communities is now possible [22]. Still, a majority of metagenomic studies concerning more complex communities employ a supra-organism framework, which considers the metagenome as a distinct, coherent entity [62,131]. Such an approach can reveal the genetic co-occurrence structure across a set of communities, but disregards the underlying structure found within constituent organisms. Notably, it has yet to be determined to what extent gene co-occurrence across communities' metagenomes resembles co-occurrence across individuals' genomes. Doing so will not only reveal whether the same evolutionary and ecological forces structure the genome and metagenome, but may unveil genetic signatures of community-driven interactions among co-occurring microbes.

Indeed, because every metagenome is formed of a composite of genomes, one might expect the metagenome to be constrained by similar forces. The metagenome, however, is a

weighted combination of genomes (*i.e.*, the relative abundances of each genome differs), genome structure may be obscured by uneven mixture of genomes. Conversely, gene pairs that do not occur in the same genomes may co-occur significantly across metagenomes, if the abundances of the organisms harboring them co-vary strongly. Such a situation may arise if, for example, phylogenetically distant organisms occupy the same niche [96,97], or if organisms support one-another's growth through co-operative interactions [132–134]. In the most extreme scenario, obligatory mutual cooperation would lead to perfect metagenomic co-occurrence among genes that are *never* found in the same genome [42,87].

In this chapter, I describe my work to characterize genetic co-occurrence across a wide array of globally distributed metagenomes and to compare it to co-occurrence found across prokaryotic genomes. I show that co-occurrence structure in metagenomes strongly resembles structure across genomes, with notable discrepancies. Specifically, while co-occurrence of core cellular processes is obscured by uneven admixture of genomes, environmental selection induces structure among genes involving secondary metabolism and xenobiotic processing. Together, this indicates a pattern of vertical inheritance of core cellular machinery, which is comparatively irrelevant to environmental selection; conversely, environment-interfacing functions (those that define an organism's niche) are more strongly represented by abundance across environments than distribution across the tree of life. Finally, we find a number of processes defined by strong intra-pathway co-occurrence in genomes, but strong inter-pathway co-occurrence in metagenomes. This supports the 'black queen' model of community assembly [133]. Energetically costly metabolic functions which benefit the community at large will be performed in their entirety by an extreme minority of benefactor species. Nonetheless, should

persistence in a given environment require numerous such functions, each function can be performed by a different community member.

4.3. Results

4.3.1. Determination of genome and metagenome co-occurrence

I collected genome profiles for all sequenced bacterial strains from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [81]. I consider only one genome from each species, and from these genomes, I filtered out invariant genes (those appearing in too few or too many genomes; Methods). Similarly, I collected shotgun metagenome profiles of a diverse array of environments from the Integrated Microbial Genomes with Microbiome Samples database (IMG/M) [135,136]. Only metagenomes with genes that passed genome-filtering were considered. In total, I generated gene-occurrence profiles describing the copy number of 4,377 genes across 1,275 genomes and 1,923 metagenomes.

To define the compositional structure of genomes and metagenomes, I calculated the co-occurrence of all pairs of genes. Because gene profiles are sparse (35% sparsity in metagenomes, 75% in genomes), correlation-based metrics (e.g., Pearson's *r*) may be inflated for pairs of genes which are absent from many of the same genomes or samples, while providing marginal information about composition. For this reason, alternative compositional metrics have been developed to study compositional structure and biogeography. For the results presented below, I utilized the cosine similarity metric: the cosine of the angle formed by two the vectors representing a pair of observations. Notably, this is a measure of orientation similarity that does not increase with shared absences.

4.3.2. Comparison of metagenome and genome compositional structure

Because every metagenome is a linear combination of individual genomes, one might expect genomes and metagenomes to exhibit similar compositional structure. Indeed, co-occurrence scores were strongly correlated between metagenomes and genomes ($r < 0.6448$, $p < 10^{-3}$, Mantel test of Pearson correlation; Figure 4.1). Additionally, as has been observed across genomes [43], and as may be expected across environments (see for example section 3.3.1 and [26]), functionally related genes co-occur more strongly in metagenomes than do unrelated genes ($p < 10^{-300}$, Wilcoxon rank-sum test; Methods) (Figure 4.2). Notably, due to the extreme sparsity of gene profiles, 8.9% of gene-pairs never co-occur in even a single genome. By contrast, only 8.8×10^{-3}% of pairs do not co-occur in metagenomes, as may be expected due the composite nature of the metagenome.

Recently the maximum relatedness subnetwork (MRS) framework has been introduced as a means to characterize global trends in genetic co-occurrence [43]. I generated networks from the genome and metagenome profiles by connecting each gene to its most strongly co-occurring partner; these networks are coarse grain representations of genome and metagenome compositional structure. Both networks decomposed into numerous discrete connected components (metagenomes: 579; genomes: 637), the sizes of which fit power law distributions (Figure 8.1). In the genome MRS, 462 components had more than one functionally categorized gene. Of these, 302 (65.37%; $p < 10^{-3}$) were uniform (composed entirely of genes of the same category; Methods), on par with previous observations. By comparison, the metagenome network exhibited a lower degree of uniformity (191 of 389 components, 49.10%; $p < 10^{-3}$), indicating an appreciable reduction in functional coherence in metagenomes

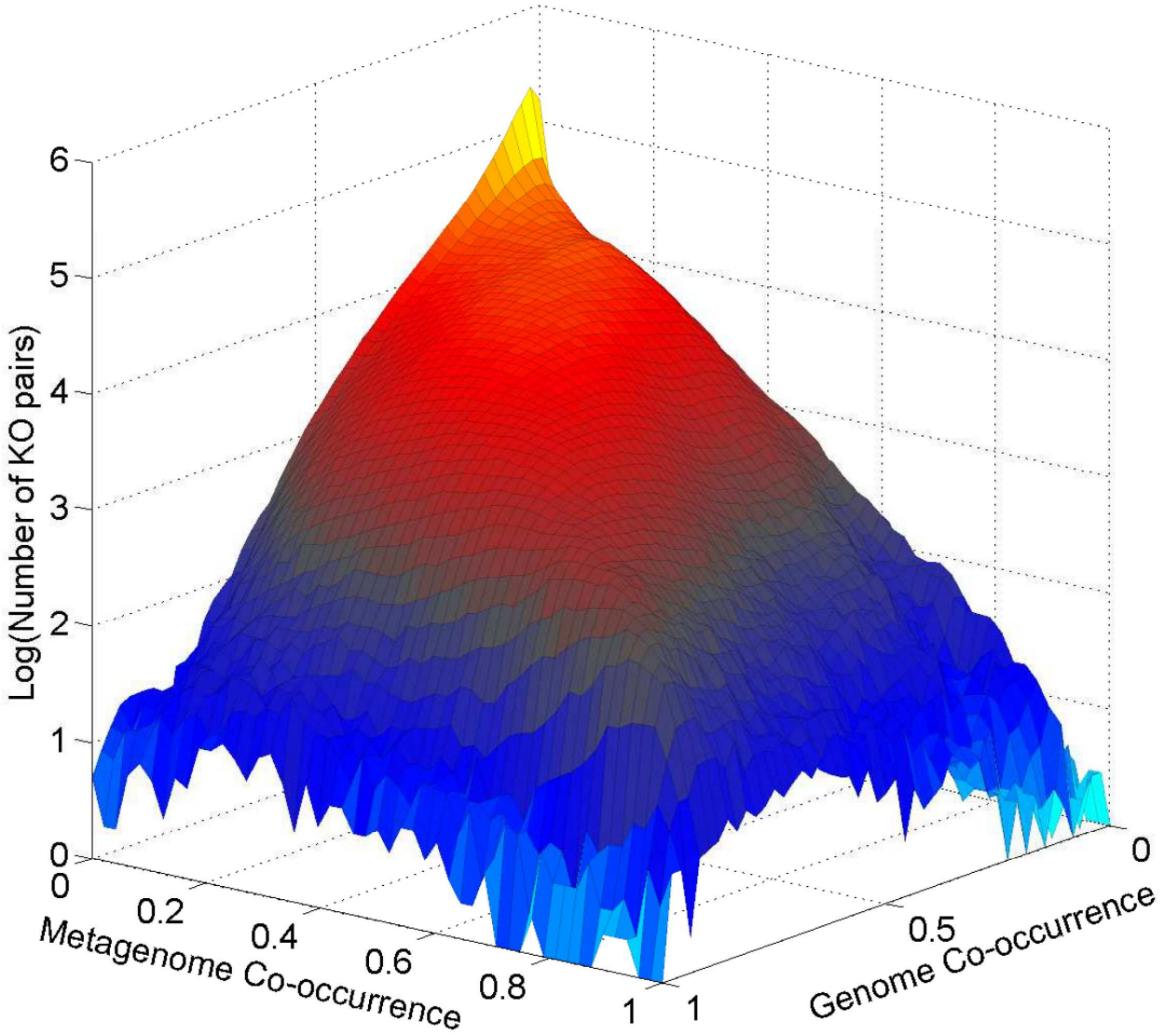


Figure 4.1: Co-occurrence of genes in metagenomes correlates with co-occurrence in genomes

The height and color of each cell scales with the log-density of gene-pairs within it. Most gene-pairs have low- to mid-range co-occurrence in both genomes and metagenomes, forming a ridge with a peak near the origin ($r < 0.6448$, $p < 10^{-3}$; Pearson correlation).

compared to genomes. Similarly, in both networks, genes of the same category tended to exhibit significant overlap (*i.e.*, fraction of genes that appear in the same component; Methods), but in metagenomes only 14 categories overlapped significantly, in contrast to 17 in genomes

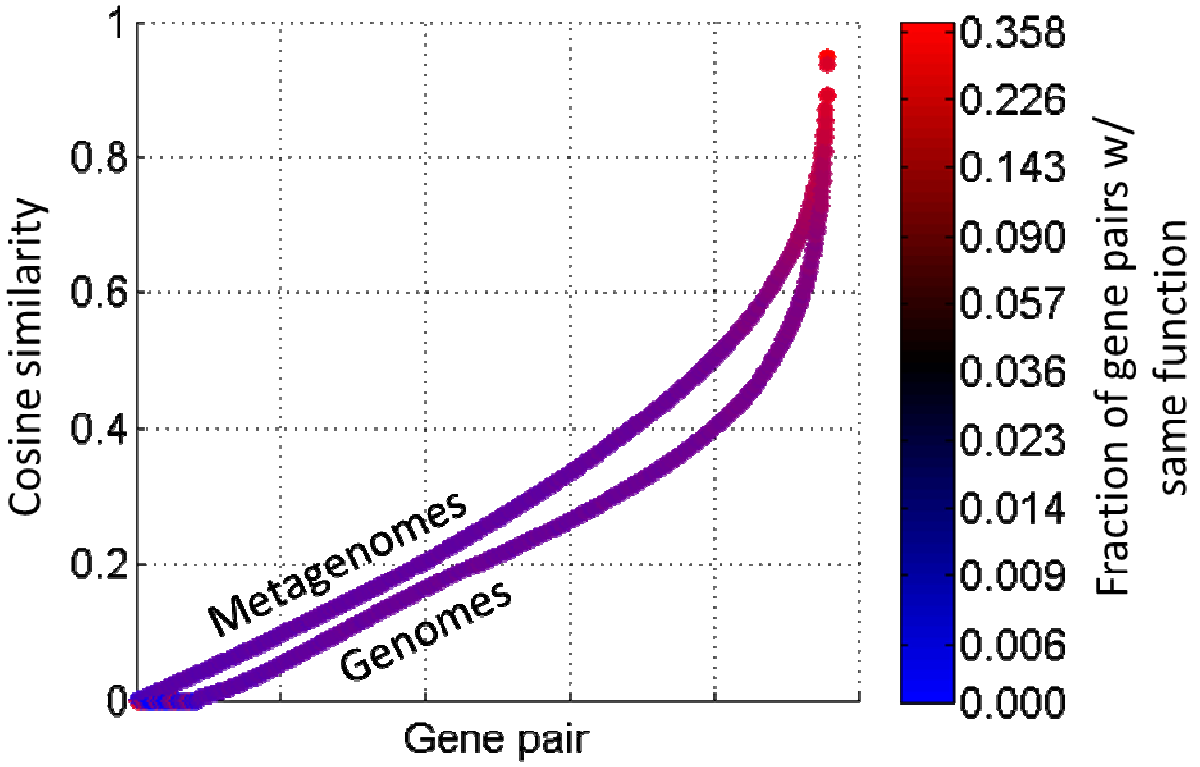


Figure 4.2: Functionally associated genes co-occur more in metagenomes and in genomes

Shown is the co-occurrence (cosine similarity) of all gene pairs across metagenomes and genomes, in rank order. Color represents functional coherence, calculated as the fraction of pairs with the same functional category within non-overlapping windows, with blue representing no and red representing the most gene pairs in a window. Window size was 10,000 gene pairs.

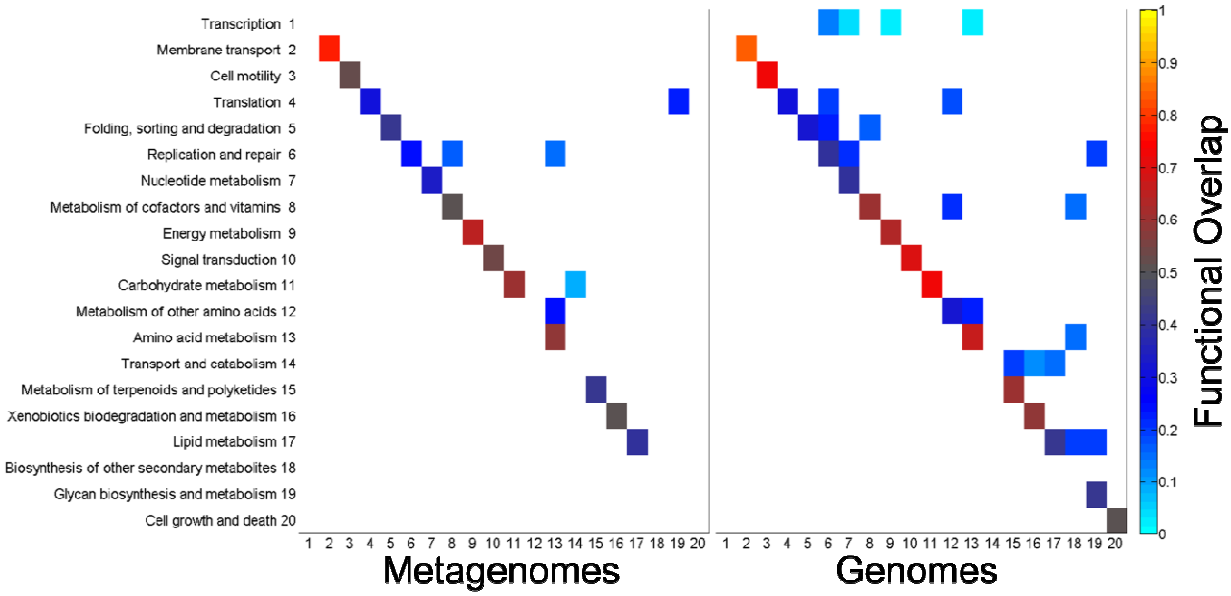


Figure 4.3: Heatmap of genetic overlap of functional categories in MRSs

The color of each element represents how likely genes of a pair of categories are to appear in the same component. Cooler colors represent low overlap, warmer colors represent increased overlap. Only significant values are shown ($p < 0.05$; Methods), and intermediate values are omitted for clarity. In both networks, Genes of the same category are more likely to overlap, but a number of off-diagonal overlap scores are significant as well.

(Figure 4.3).These results are in agreement with the hypothesis that through admixture, structure appearing in genomes will appear reduced in metagenomes.

Some pairs of categories also overlapped significantly one another, although fewer of these 'off-diagonal' category pairs were found in metagenomes than in genomes (5 and 19, respectively). Notably, many function pairs with significant overlap in genomes but not in metagenomes related primarily to core cellular processes (e.g., *translation* with *replication and repair*), suggesting that while these processes are tightly coordinated in single organisms, this association is not maintained at the community level. In contrast, pairs significantly overlapping in metagenomes tended to link core processes to secondary metabolic functions (e.g., *translation* with *glycan biosynthesis and metabolism*). Finally, the only pair of categories that overlapped significantly in both the genome and metagenome MRSs the biosynthetic pathways

of either proteinogenic or non-protenogenic amino acids (*Amino acid metabolism* and *metabolism of other amino acids*, respectively. Pathways of these categories are metabolically linked (e.g., taurine is reversibly transformed to alanine in anaerobic organisms such as *Bilophila wadsworthia* [137], and may represent cooperative exchanges between co-occurring auxotrophs [42]; it is unsurprising to find these functions coordinated at the genome as well as metagenome level. In all, the observed differences in MRS structure suggest that while the strongest compositional features of genomes are preserved in metagenomes (*i.e.,* co-occurrence of functionally related genes) discrepancies in structure may indicate functions coordinated across the community level.

4.3.3. Differential analysis of co-occurrence structure distinguishes assembly of genomes and metagenomes

I next characterized the difference in compositional structure between metagenomes and genomes. To do so, I calculated the average co-occurrence of genes within each functional category, both across metagenomes and genomes. Within some categories genes co-occurred more strongly in metagenomes than in genomes, and vice-versa (Figure 4.4, diagonal elements). Specifically, genes associated with translation and cell mobility co-occur most strongly in genomes relative to metagenomes. Notably, the translational machinery (in particular the ribosome) is the prototypical indicator of phylogenetic history [138,139], and a complete suite of motility genes is all but essential for proper function [140]. Strong genetic co-occurrence within these molecular machines is likely the result of vertical co-inheritance driven by molecular interactions essential for proper function. By contrast, secondary metabolic and glycan biosynthetic genes exhibit the highest co-occurrence in metagenomes relative to genomes. These functions interface directly with the environment, and are stronger indicators of ecological

niche [54,59,96,97]. Interestingly, whereas core cellular machinery is vertically transmitted, it has been shown that ecology drives the horizontal transfer of genes which define an organism's niche [91], leading to functional convergence of phylogenetically distant taxa [26,69].

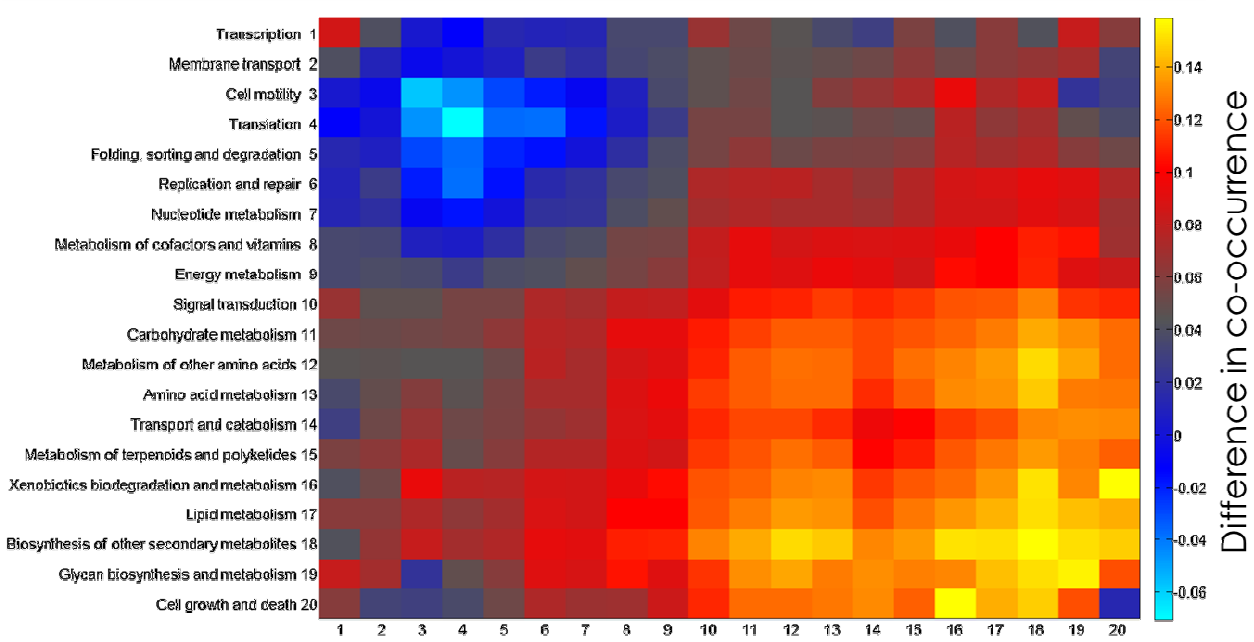


Figure 4.4: Different functions co-occur more often in metagenomes than in genomes

For each pair of 20 categories, the average gene co-occurrence of genes in metagenomes and in genomes was calculated. Shown is the difference in mean co-occurrence; hot colors indicate function pairs which, co-occur relatively strongly in metagenomes compared to genomes; cooler colors indicate function pairs which co-occur strongly in genomes. Diagonal elements represent the co-occurrence of genes of a given function category.

I next investigated which pairs of functional categories are more coordinated across metagenomes than across genomes. I found two clusters distinguished by the relative strength of between-category genetic co-occurrence in genomes and metagenomes (Figure 4.4 and Figure 8.2). As expected, the genome coordinated cluster principally contained genes associated with core cellular processes such as genetic information processing as well as central energy metabolism, while the metagenome coordinated cluster comprised genes associated with the cell-environment interface or secondary metabolism. Interestingly, despite

the fact that genes of the category *cell growth and death* are vertically co-inherited, this category clusters with metagenome coordinated processes. These genes largely involve regulation of the cell cycle, indicating environmental clustering of organisms with similar ecological strategies and growth rates [60].

4.3.4. Analysis of metagenome coordinated pathways reveals community distributed processes

In order to characterize the assembly of metagenomes at a finer resolution, I repeated the above analysis by associating genes to KEGG pathways rather than to categories. While KEGG pathways describe microbial physiology at a finer resolution, each is nonetheless defined broadly enough to encompass an array of specific molecular functions. Surprisingly, genes from the pathways *Phosphotransferase system* and *ABC transporters*, both involved in transport of environmental compounds, co-occur across genomes more than metagenomes despite being strongly related to habitat preference and varying greatly across metagenomes (see also chapter 3). In contrast, of the 15 pathways with the greatest co-occurrence in metagenomes relative to genomes, many involve aromatic hydrocarbon metabolism. Furthermore, 11 of these were degradation pathways, 5 of which specifically involved *xenobiotics biodegradation*, indicating that the ability to degrade or process environmental hydrocarbons may be among the strongest factors contributing to metagenome assembly.

In order to identify process which may be distributed amongst co-occurring microbes, I searched for pathways with strong intra-pathway gene co-occurrence in genomes but with high inter-pathway coordination in metagenomes. Organisms which operate one of these pathways do not operate the others, but co-occur with those that do. I found 4 such pathways,

glycosphingolipid biosynthesis - globo series, *phenylpropanoid biosynthesis*, s*esquiterpenoid and triterpenoid biosynthesis*, and *fluorobenzoate degradation*. Fluorobenzoate is a xenobiotic resistant to microbial degradation [141] while the pathways are commonly used by plants as antimicrobials (see also Discussion). Conversely, *steroid degradation* genes co-occur with one another as well as with *steroid hormone biosynthesis* genes strongly in genomes, yet genes within the steroid biosynthesis genes co-occur with one another more strongly in metagenomes. This may potentially indicate that particular steroid hormones synthesized by distinct community members, or perhaps even that steroid metabolism is distributed among community members [142].

4.4. Discussion

As shotgun metagenomics becomes the method of choice for characterizing microbial community structure, the supraorganism framework will become the paradigm for modern microbial ecology. This framework considers the metagenome a distinct entity that represents the activity of a community *en masse*, the composition of which varies in response to environmental pressure. Consequently, as the diversity of a community is defined not only by the number of species within it but also by their relative abundances, so too is metagenome structure defined by more than the complement of genes within it, but the relative abundance of each. And, as each gene's abundance is contributed to by a diversity of species' genomes, the compositional structure of a metagenome is not unconditionally constrained by the taxonomic structure of the community it represents [2].

Accordingly, the structure of a metagenome does not innately mirror the structure of the genomes which compose it, and discrepancies in genome and metagenome structure may

reveal features of the community assembly process. Given such considerations, the observation that the most fundamental aspect of genome structure, gene co-occurrence among functionally related genes which may not directly interact, is maintained within the metagenome supports the supraorganism paradigm. Specifically, it implies that community members respond to environmental pressures in coordinated manners. Were each species to employ a distinct coping strategy, intra-pathway co-occurrence would be greatly reduced among metagenomes.

This phenomenon is in fact observed within some cellular activities. Specifically, within core cellular processes such as genetic information processing, intra-pathway co-occurrence is reduced in metagenomes compared to genomes. This likely represents the relative irrelevance of these processes to habitat selection; the genes encoding these functions are typically vertically co-inherited. Interestingly, environments which heavily influence the choice of information processing machinery (e.g., hot springs) tend to be associated by a narrowly defined taxonomic range (e.g., dominated by the kingdom archaea), supporting strong phylogenetic signatures of vertical inheritance. Conversely, processes which define an organism's niche may be more coordinated at the metagenome community, as these genes specifically drive the relative abundance of their hosts across environments. Indeed, we observed such pathways are more coordinated in the metagenome than genome (Figure 4.2).

Finally, metagenome assembly takes into account contributions of all community members. The recently postulated *Black Queen Hypothesis* supposes that within an environment, metabolically costly processes which support the community at large will be operated by a specific, narrow subset of the community [133]. Critically, in environments necessitating many such functions, only one species needs to provide each. Indeed I found that

the degradation of xenobiotics, extremely costly functions benefiting all neighbors, is coordinated at the metagenome level. Interestingly, this analysis also revealed xenobiotic and phytochemical degradation pathways with strong phylogenetic signatures that are nonetheless cross-coordinated in metagenomes, potentially indicating the distribution of each such pathway across specialized community members; each microbe plays its own role in community assembly and preservation.

A number of considerations need to be discussed in interpreting these results. In any such study, the granularity with which genes are characterized is critical. In a manner analogous to species' phylogeny, gene products are functionally defined along a hierarchical tree, ranging from specific enzymatic activity through pathways and, most broadly, to cellular processes. While this study and others demonstrated strong constraints on the metagenome at the functional level [2], these were observed at broad functional definitions. Just as ecological processes play out different at different taxonomic scales [35,109], it remains to be seen to what degree these communities exhibit significant functional differences at a finer resolution.

Perhaps the most critically is the fact that the vast majority of genes lack *any* functional annotation [3,123]. I considered here only the subset of genes with functional annotation, the vast majority of which were inferred solely through homology to characterized genes. Similarly, an extreme minority of species on the tree of life has been sequenced, and significant bias exists in which clades these come from [122,143]. As most environmentally and ecologically interesting genes and genomes may be found in this 'dark matter,' their role in community assembly is yet unknown. Nonetheless, the analysis of genetic co-occurrence employed here has been suggested as a means to begin to implicate by association genes of unknown function

[43]. The use of metagenome information may provide an additional layer of information, as differential co-occurrence implies specific cellular processes for genes, indicating yet another intriguing means by which microbial ecology can contribute to our understanding of microbial life at multiple scales.

4.5. Methods

4.5.1. Genetic co-occurrence in metagenomes and genomes

Gene content for all bacterial species was collected from the Kyoto Encyclopedia of Genes and Genomes (KEGG, [81]). Organisms listed as *draft* or *new* were discarded. Subsequent data pre-processing followed the example of [43]. To reduce bias from uneven representation among bacterial clades, only the earliest sequenced genome from each species was selected. Genes in KEGG are organized into KEGG orthologous groups (KOs), which are referred to simply as genes throughout the chapter. Invariant genes were removed using the method previously described: in this study, genes present in (absent from) fewer (more) than 40 genomes were discarded. For the remainder, the copy number of each gene in each genome was recorded. Metagenome data was collected from the Integrated Microbial Genomes with Metagenomes database (IMG/M) [135,136]. KO abundance profiles (estimated gene copy numbers) for all finished, draft, and permanent draft metagenomes were downloaded. Synthetic and mock communities were removed by manual inspection. The data were normalized using MUSiCC [144], to correct for intra- and inter-sample variation.

As described, the resultant metagenome and genome gene profiles are very sparse. Therefore, many pairs of genes will be absent from the same genomes and metagenomes. In these cases gene-pairs may actually appear to be highly correlated even if they strongly

segregate in the subset of samples in which at least one appears. For this reason, ecological analyses of species or trait data typically utilize set-based similarity metrics which are unaffected by co-absences. To define the similarity of gene-incidence profiles, I utilized the cosine similarity, which can be described intuitively as the cosine of the angle formed by two observations when represented as vectors in feature space:

$$S_{cos}(A,B) = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

Notably, since gene copy number is positive, the cosine similarity is bounded [0, 1], with 0 meaning complete segregation, and 1 meaning complete overlap (one gene profile is a scalar multiple of another).

4.5.2. Functional categorization of genes

I annotated genes with function definitions at two resolutions: KEGG categories, which represent a broad definition (e.g., *energy metabolism*), and KEGG pathways, which define a more specific metabolic or cellular function (e.g., *oxidative phosphorylation*). Genes may be annotated to more than one such category or pathway. Any category or pathway which contained fewer than 3 genes in our dataset were discarded, as were functions not performed by microbial species (e.g., *human diseases*, *Huntington's disease*), resulting in 20 category and 128 pathway terms.

4.6. Maximum relatedness subnetwork (MRS) analysis

Separate MRSs for metagenomes and genomes were generated from gene-gene co-occurrence profiles. Nodes represent genes, and a directed edge is drawn connecting each

gene to its most strongly co-occurring (or excluding) partner(s). In the event of ties, edges connect a source to all equally co-occurring (excluding) partners. Considering only co-occurrence edges, both MRSs decomposed into numerous distinct connected components. Because co-occurrence as defined here is bound [0, 1], the criterion that exclusion edges be negative was ignored. Notably, this resulted in a greater density of exclusion edges, as many genes completely segregate, and all such gene pairs are connect by exclusion edges. Nonetheless, I find that of 1,704,708 exclusion edges in genomes (5,727 in metagenomes), only 20 such edges connect genes of the same component (2 in metagenomes). Functional coherence and overlap were calculated using methods described previously in Ref. [43].

5. Concluding remarks and future considerations

In this dissertation I have addressed the question of structure in the microbiome. Specifically, I demonstrated the existence of community assembly rules, deterministic processes which produce the structure microbial communities exhibit across habitats. In chapter two I demonstrated that the microbiome (particularly the human microbiome) is structured predominantly by habitat filtering rather than by species assortment. In chapter three I demonstrated that co-occurring microbes partition their niches rather than seek out cooperative interactions. Finally, in chapter four I demonstrated that the community metagenome has structure not contained in individual genomes which highlights processes coordinated across the community as a whole. As I discussed, elucidating these assembly rules required novel analytical frameworks integrating metagenome and whole genome sequence information. These approaches are now being referred to as the **metagenomic systems biology** framework, and will continue to provide an extensive methodology to investigate the structure and function of microbial communities. In this chapter I discuss the interplay of community assembly rules in structuring the microbiome, as well as provide perspective on the application and development of the metagenomic systems biology framework.

5.1. A Microbial hierarchy of needs

The assembly of a microbial community is a complex process demanding stable equilibrium between antithetical ecological forces. The relative contribution of each ecological process leads to what may be represented by a microbial hierarchy of needs (Figure 5.1). The

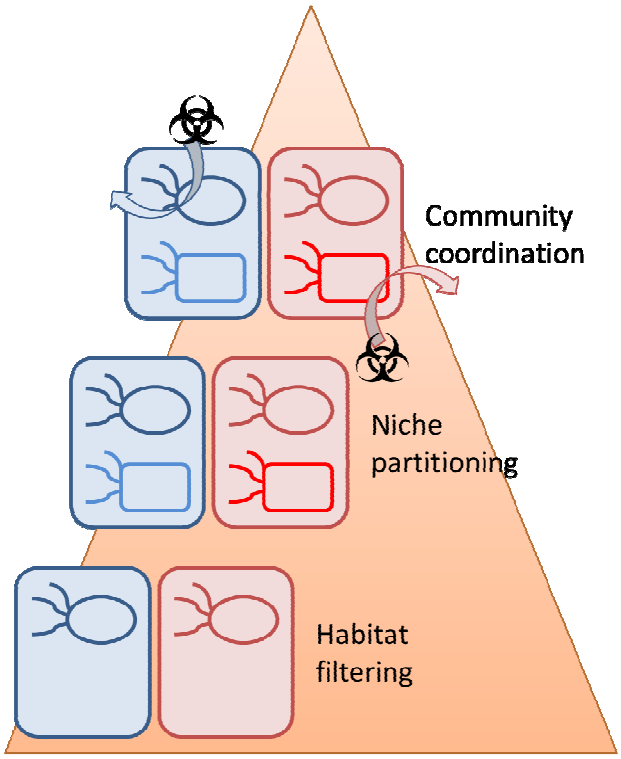


Figure 5.1: A microbial hierarchy of needs

This diagram represents the relative roles of three key processes defining the assembly of microbial communities. The most basic community assembly rule is habitat filtering, by which organisms are distributed only to those environments most favorable to their growth (represented by color coordinated assignment of microbes to habitats). Subsequently, niche partitioning limits the negative effects of interspecific competition among cohabiting microbes, while still constrained by habitat choice (represented here by microbe shape and shading). Finally, once these considerations are accounted for, community coordinated processes such as environmental remediation can be distributed among community members (represented by individual species' degradation of contaminants).

most basic consideration of a microbe is the presence of favorable environmental conditions, such as temperature, salinity, and nutrient availability. Thus, habitat filtering is the foundation of community assembly: microbes are limited to the range of habitats for which they are specifically adapted, and therefore co-occur chiefly with similarly adapted neighbors. As complete niche overlap would lead to competitive exclusion [145], community members must subsequently differentiate niches to an extent sufficient to support co-occurrence. Finally, once these objectives are met distributed processes play a role in supporting the community: those

functions costly to the individual yet beneficial to the community are provided by a subset of community members [133].

Notably, this model considers species-assortment subordinate to habitat filtering; it suggests that competitive interactions are not likely to produce pairs of species which never co-occur, as proposed by the theory of limiting similarity [145]. Instead, such interactions lead to segregation in niche space rather than physical proximity. Also, an alternative interpretation of the assembly process might place community-coordinated processes as a foundational role alongside habitat filtering. Specifically, this interpretation views the modifications one organism makes to its environment as redefining the habitat filter acting on potential neighbors. Under such a model, community assembly rules would be separately defined for each constituent member, including the existence of neighbors as environmental features. In considering the community as distinct from the environment, the model presented in this work operates within the supra-organism paradigm to provide a single description of the ecological forces defining a climax community in a unified manner.

5.2. Future directions

5.2.1. The dark space of the tree of life

While these results demonstrate the potential promise of metagenomic systems biology, this framework shares the same limitations inherent to any meta'omic analysis. Most critical of these is the need proper functional characterization of genes. As many as 75% of genes identified in the human metagenome lack any functional characterization [3], and the fraction of uncharacterizable genes is likely higher in other environments. These genes have been called the *functional dark matter* of the microbiome, analogous to the *phylogenetic dark matter*

previously described [122,123]. The ideal solution to this problem is the meticulous biochemical characterization of these gene products, but the organisms harboring them tend to resist culture and isolation [31]. Individual genes sequences determined by shotgun sequencing can be cloned into model organisms, but even in this scenario it is unlikely that the product will be produced efficiently and behave in a novel genomic context as it does in nature.

5.2.2. Better models of metabolism

This work made exclusive use of topological models of metabolism. These models can be applied to nearly all organisms associated with a whole genome sequence, and metrics describing global structural properties are robust to annotation error [54]. Nevertheless, as they do not encode reaction stoichiometry, they cannot be used to describe metabolic activity (e.g., enzyme kinetics or fluxes). A number of constraint based models of community metabolism have been introduced [14,74,90,94,95,146,147], each of which assumes a different community-level objective. In the same manner that biological intuition (of evolutionary and regulatory dynamics) guided the choice of optimal single species objectives [148,149], ecological intuition gained using methods presented in this work may guide future models of community metabolism. For example, a conceivable additional constraint on the community may come in the form of maximizing the overall coverage of niche space, or minimizing the utilization of shared metabolic inputs.

5.2.3. Integration of more data

This work established the value of integrating multiple 'omic datatypes to analyze microbial communities, yet it did not exploit the full range of available data. Recent studies have collected metagenomic, metatranscriptomic, and metametabolomic data of the same

communities. Such studies performed on single species have, for example, revealed surprising stability of metabolic activity achieved through tight regulation of alternative transcriptional programs [150]. Similar approaches may finally explain the apparent similarity of microbial communities at the functional level despite significant taxonomic variation [2].

5.2.4. Synthetic ecology

Genome scale computational models provide powerful tools to generate hypotheses of microbial physiology, which are relatively easily tested using experimental synthetic biology approaches. Computational and experimental synthetic ecology methods have been introduced as the analog to these approaches at the community scale. Nonetheless, owing to the increased scale and complexity of community compared to organism systems, our understanding is relatively lacking. Bridging the gap between these approaches and meta'omic analyses of natural communities provides promising opportunities, but also critical challenges. As described in chapter 3, these approaches may appear to produce contradictory results. In such scenarios careful consideration of the limitations, assumptions, and methods used by both approaches are essential to explain supposed discrepancies. Similar to classical colonization studies which integrated large scale data with community perturbation experimentation performed directly in the field [151], natural communities can be perturbed and their response studied analyzed using 'omic techniques. Notably, invasion experiments using genetically modified microbes performed on a small scale on synthetic communities [11], highlighted specific genetic factors modulating ecological processes, demonstrating an advantage such approaches have over even their macroscopic equivalents.

5.2.5. Intervention

The goal of metagenomic systems biology is the development of a predictive model of community behavior [58]. Such a tool would not only demonstrate a comprehensive understanding of microbial communities, it would provide a foundation for targeted microbiome intervention or even design of novel microbiomes. In this regard, most attention has surrounded fecal microbiota transplantation: a donor intestinal microbiota is transplanted to an ill patient, perhaps suffering from acute *clostridium difficile* infection or IBD. While markedly successful, no model has been demonstrated to explain the efficacy or stability of the transplanted community. Answering these questions will likely require models of species interactions like those presented in this work.

Another potential application lies in the targeting of specific community members, for example to culture fastidious organisms. Some organisms may not have been successfully cultured because they require the presence of a partner species. Models of community metabolism may identify such interactions, allowing for the co-culture of such consortia. Another application of targeted intervention may be the identification of species (probiotics) or metabolites (prebiotics) which specifically effect the growth of clinically relevant organisms. Such an approach could, for example, protect against the expansion of pathogens after antibiotic treatment, or increase the production of microbial products beneficial to a host.

Clearly much work still remains before we have a comprehensive understanding of the microbiome. The description of community assembly rules represents merely the first step of an ongoing process. The long term objective of metagenomic systems biology, the generation of predictive models of microbiome composition to aid in the design of microbiomes to meet

specific needs, is clearly an ambitious goal. Nonetheless, due to their importance to human and to ecosystem health, the benefit of such capabilities is clear. Fortunately, the progress achieved using metagenomic systems biology, by these as well as numerous other studies, justifies an optimistic outlook moving forward.

6. Appendix A: Supporting Information for Chapter 2

6.1. Supporting text

6.1.1. Human oral community stringent growth comparison

I performed a more stringent analysis of the human oral community, repeating the growth rate comparison described above (Methods) but using my own assessment of growth. Specifically, in this more stringent analysis, I used manually curated data gathered from previous studies [65,66,152–155] and compared only growth rates measured as part of the same experiment (*i.e.*, the same experimental conditions and in the same reference) and without conflicting evidence (e.g., if two species grow on pegs but not in flow-cells). This more stringent analysis provided qualitatively similar results. Species with better impact on the growth of a partner tend to have lower metabolic competition indices and higher metabolic complementarity indices ($p < 0.048$ and $p < 0.024$, respectively; paired one sided t-test).

I additionally confirmed that *P. gingivalis'* metabolic competition and complementarity with other community members represents its ability to form mutualistic biofilms with many species [66]. Specifically, I demonstrated that *P. gingivalis* is the most complemented by and poses the least competition to all other species. I first noted that for 3 of the 6 target species it poses the lowest metabolic competition and in the other 3 the second lowest (Table 6.2; see Pg column); in all 6 cases it is the most complemented species (Table 6.3; see Pg row). Comparing interaction indices associated with *P. gingivalis* to those associated with other species, I additionally found that the set of scores denoting the metabolic competition posed by *P.*

gingivalis are significantly lower than all other corresponding scores ($p < 0.003$, one-tailed Wilcoxon rank-sum test). Similarly, the scores denoting the complementation received by *P. gingivalis* are significantly higher than all other complementarity scores ($p < 1.28 \times 10^{-4}$, one-tailed Wilcoxon rank-sum test). Finally, examining all pairwise comparisons of the competition scores posed by *P. gingivalis* (or the complementarity scores received by *P. gingivalis*) to that of other species using a one sided rank-sum test, I found that the median competition associated with *P. gingivalis* is lower in all 6 comparisons, significantly so in 4 of 6 comparisons ($p < 0.01$, one-sided Wilcoxon rank-sum test); the complementarity scores associated with *P. gingivalis* are significantly lower in all cases ($p < 0.01$, one-sided Wilcoxon rank-sum test). These results are in line with observation of the interaction between *P. gingivalis* and other oral species; while it is not necessarily the preferred growth partner of all species, it can form mutualistic biofilms with many species.

6.1.2. Alternative co-occurrence metrics and sensitivity to under-sampling

For the results reported in the main text I applied the widely used Jaccard similarity index as a measure of co-occurrence [156]. However, since a number of co-occurrence metrics have been used in previous ecological studies (e.g., Refs [157,158]) and no standardized similarity metric has been fully established, I confirmed that my main findings are not an artifact of the specific metric used and that the observed patterns hold under several alternative measures of co-occurrence. Specifically, I examined the correlation between my interaction indices and several previously introduced ecological co-occurrence metrics, including in addition to the Jaccard similarity index, the Bray-Curtis similarity, the Morisita-Horn similarity, and Cosine similarity. I additionally repeated my analysis using a range-normalized transform of the abundance data: for each species, the abundance was scaled such that the lowest observed

abundance value was 0 and the highest was 1. In this way, these metrics are not quantifying the similarity of abundance profiles, but rather the similarity in the changes of abundance across samples. For example, without such normalization, species that have very high abundances in all samples would appear to have similar profiles, even when a rise in the abundance of one is associated with a decrease in the abundance of the other. I found that using any of the above metrics or normalization schemes did not qualitatively change the results reported in the main text (See Table 6.13). I additionally examined Pearson and Spearman coefficients of correlation as a measure of co-occurrence. Using these measures, which are known to attribute spurious associations in relative compositional abundance data [159], resulted in generally similar, but weaker patterns.

Due to the limited number of individuals sampled by the MetaHIT study, I further sought to determine whether under-sampling of individuals might have any detrimental effects on observed co-occurrence values. I repeatedly subsampled at random 62 individuals (50% of total) uniformly, with no regard to nationality, health state, BMI or enterotype. Using these samples, I re-calculated the co-occurrence of all species pairs using all metrics described above. I found Jaccard similarity index to be the most robust, with relatively little variation in obtained co-occurrence values (Figure 6.1). Consequently, I used this co-occurrence metric to report the results of my analyses.

6.1.3. Analysis of coherently predicted interactions

Since the nutritional profiles of species vary substantially in size, my predicted interaction indices are not necessarily symmetric. Consider, for example, a species A with a nutritional profile containing many compounds, and a second species B with a nutritional profile

containing only a few compounds, all of which also appear in the nutritional profile of species A. In this extreme example, the metabolic competition index of species A on species B is 1 while the metabolic competition index of species B on species A is much smaller than 1 (and approaches 0 as the size of A's nutritional profile increases). Since it is hard to interpret the exact effect of niche overlap in such extreme scenarios (see, for example, [17]), I sought to control for these cases. I therefore repeated my analysis using only coherent interactions: pairs of mutual indices that are within 0.1 of one another. I found that with this control the observed association between predicted interaction indices and co-occurrence increases and is still highly significant (ρ = 0.249 , $p < 10^{-4}$ & ρ = -0.204, $p < 10^{-4}$, metabolic competition index & metabolic complementarity index respectively, Mantel correlation test). I additionally found that as the definition of coherent interactions is made more stringent, the magnitude of correlation between interaction indices and co-occurrence increases, potentially indicating that strongly reciprocated interactions exert a larger influence on co-occurrence patterns. I similarly reevaluated these results analyzing only pairs of species with nutritional profiles whose sizes are within 10 compounds. Again, I found a similar association (ρ = 0.285, $p < 10^{-4}$ & ρ = -0.193, $p < 10^{-4}$, metabolic competition index & metabolic complementarity index respectively, Mantel correlation test).

6.1.4. Comparison of species' partners and excluders to a null model

I used the Mantel Test to compare the interaction indices of partners and excluders to a null distribution. Each species' partners and excluders were determined as described before. The number of species that have greater metabolic competition with partners than with excluders, and lower metabolic complementarity with partners than with excluders was determined. To determine the significance of these associations, I randomly shuffled species

co-occurrence 10,000 times. For each shuffled matrix, I again determined partners and excluders, and their mean metabolic indices. The *p*-value was calculated as the fraction of shuffled matrices in which a higher or equal number of species were observed with greater metabolic competition with partners than with excluders, or with lower metabolic complementarity with partners than with excluders. I found that the separation of partners and excluders by metabolic competition index and metabolic complementarity index was significantly high ($p < 2x10^{-4}$ & $p < 1x10^{-4}$, respectively, Mantel test).

6.1.5. Metabolic competition of consistent and inconsistent partners and excluders

I determined the consistency of co-occurrence patterns across health states and examined whether consistency is associated with predicted metabolic interactions scores. To this end, I partitioned the samples into two groups: healthy individuals and those with IBD. I determined each species' partners and excluders in each group separately as before (Methods). I found that the vast majority (96%) of co-occurrence partnerships were consistent across health states: 3800 pairs consistently co-occur and 3912 pairs consistently exclude across healthy and IBD samples, while only 140 pairs co-occur in healthy and exclude in IBD and 201 pairs exclude in healthy and co-occur in IBD. Examining the metabolic competition index for consistent and inconsistent species pairs separately, I found a significant association between co-occurrence and metabolic interaction: consistent partners exhibit higher metabolic competition than consistent excluders, with inconsistent partners/excluders exhibiting intermediate competition levels (Figure 6.2).

6.1.6. Consistency of species' partners and excluders separation across species' ecological traits

I examined whether the difference observed between a species' partners and excluders is consistent across species and specifically whether species associated with a certain ecological attribute escape these assembly rules. To this end, I collected species' ecological characteristics from the NCBI Genome Projects' table of prokaryotic genomes (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi). Six species did not have an exact strain match, and an alternative strain was used. Three characteristics relevant to the ecology of species were recorded: oxygen requirement, habitat preference, and pathogenicity. These data were used to label each species with a subset of the following attributes: pathogen, human pathogen, anaerobe, facultative anaerobe, host-associated, and cosmopolitan (Table 6.6). Species not listed as pathogenic were assumed to be benign. Other omitted annotations were treated as missing information. For each ecological attribute (e.g., anaerobes), the number of species for which partners have higher mean competition index than excluders or for which excluders have higher mean complementarity index than partners was counted, as well as the number of total species for which information about this ecological attribute is available. I used a hypergeometric enrichment test to determine whether any of the ecological attributes tested is enriched among species with properly separated partners and excluders. I found that species labeled with each of these attributes exhibit a similar pattern in terms of their metabolic competition and complementarity with partners and with excluders as those not labeled with the attribute (Table 6.14).

6.1.7. Metabolic versatility cannot fully account for the observed habitat filtering patterns

Species with larger nutritional profiles (*i.e.*, larger seed sets) are potentially more metabolically versatile and may be able to survive in many environments using subsets of their nutritional profiles [54,59]. Such environmental generalists may be therefore able to survive with a wider range of interacting species, mitigating the competitive influence of niche overlap. I therefore confirmed that the differences in metabolic interaction indices between partners and excluders were not in fact the outcome of variation in nutritional profile size. First, I found no significant correlation between nutritional profile size and mean co-occurrence rank (ρ = 0.049, p < 0.270 Mantel Correlation Test), suggesting that such species will not necessarily be considered partners of many other species. Furthermore, I calculated the partial correlation between co-occurrence and metabolic interaction indices and found that controlling for nutritional profile size does not lead to a significant reduction in correlation (ρ = 0.210, $p < 0.10^{-3}$ & ρ = -0.199, $p < 0.10^{-3}$, metabolic competition and complementarity, respectively; Mantel partial correlation test). I also examined whether the nutritional profile size of partners is consistently different from that of excluders at different phylogenetic distances, and exhibits a similar pattern to that observed for metabolic competition in Figure 6.1. Using the same phylogenetic relatedness bins as those used in Figure 6.1 I found that the difference in nutritional profile size between partners and excluders is not consistent. In only 4 of the 6 bins do partners have larger nutritional profiles than excluders, and in the most populated bin partners in fact have significantly smaller nutritional profiles. Finally, determining the average nutritional profile size among the partners and excluders of each species I found that for only 58% of the species (90 of 154; a non-significant enrichment) do partners have larger nutritional profiles than excluders,

compared to 82% of the species in which partners have higher metabolic competition than excluders.

6.1.8. Comparison of competition, complementarity, and phylogeny in distinguishing partners vs. excluders

I examined the ability of three different indices in distinguishing between each species' partners and excluders: the metabolic competition index, the metabolic complementarity index, and phylogenetic relatedness determined by 16s similarity. I compared these metrics in partners and excluders of each of the 143 species for which estimates of phylogenetic relatedness are available (Methods). As before, I classified as partners of a given species the 25% of species with which it has the highest co-occurrence scores, and as excluders the 25% of species with which it has the lowest co-occurrence scores. I found that each of the three indices above distinguishes partners and excluders roughly equivalently: 81% of species (116 out of 143) have greater metabolic competition with partners than with excluders, 86% (123 out of 143) have lower metabolic complementarity with partners than with excluders, and 78% (112 out of 143) have greater phylogenetic relatedness with partners than with excluders. I found, however, that the sets of species for which each index correctly distinguishes partners and excluders is not identical (Figure 6.3), suggesting that these three criteria, competition, complementarity, and phylogeny, encapsulate distinct information about the co-occurrence of species.

6.1.9. Testing host nationality and enterotype

Since the data included samples from two different cohorts, Danish and Spanish, I examined whether variation between these two cohorts can account for the observed habitat filtering pattern. Partitioning our samples and repeating the analysis above considering

separately samples from each nationality, I did not find any qualitative change in the trends reported (Table 6.10). Furthermore, since it has recently been suggested that variation in the human intestinal microbiota tends to cluster into three discrete states (termed enterotypes [160]), I confirmed that the association between co-occurrence and metabolic interaction indices holds when controlling for enterotypes found in our dataset (Table 6.10).

6.1.10. Correlation of co-occurrence and metabolic interaction indices in HMP oral samples

In the oral community, the observed correlation between co-occurrence and metabolic interaction was found to be generally weaker than the correlation obtained for other body sites and was not statistically significant for the metabolic complementarity index. The lack of a clear habitat filtering signature within the oral community may be attributed to a number of factors. First, whereas the other body sites generally represent a single specific subsite (e.g., the nares in the airways), the oral community was sampled from several distinct subsites, each of which represents a specific niche [68]. Second, while the α-diversity of the oral community is higher than other communities [2], the number of organisms surveyed that mapped to sequenced genomes was similar to other sites, potentially underrepresenting the community and making it more susceptible to the influence of noise in the data.

6.1.11. Alternative reverse ecology interaction index

In addition to the interaction indices discussed in the main text, I also investigated the association of species co-occurrence with a previously described reverse-ecology interaction measure, the Effective Metabolic Overlap (EMO) score [161]. Similarly to the metabolic competition index I developed, EMO is a network-based algorithm for estimating the competition

between two species. The two indices, EMO and metabolic competition index, are significantly correlated ($\rho = 0.312$, $p < 10^{-3}$). It is important however to note a fundamental difference between these two measures: while the metabolic competition index directly quantifies the amount of niche overlap between species, EMO aims to quantify the deleterious downstream effects of a competing partner on the growth of a species. Briefly, to determine the EMO of two species, the nutritional profiles of both species are calculated, overlapping metabolites are removed from the nutritional profile of the query species, and the network expansion algorithm [162] is performed to determine how many essential metabolites this species is still capable of synthesizing. Using the EMO score to predict competitive interaction between species, I obtained qualitatively similar results to those observed using our metabolic competition index: EMO correlated positively with co-occurrence (correlation was weak but significant). Restricting our analysis to coherent EMO scores further improved the correlation ($\rho = 0.140$, $p < 10^{-4}$). In 110 species (71%), partners have greater EMO than excluders.

6.1.12. Consistency in definition of partners and excluders

I tested the robustness of my results to the definition of partner and excluder species. I defined each species' partners as those species with which it shared the greatest co-occurrence, and excluders as those with which it shared the lowest. In the main text, I used a threshold of 25% of species for determining high and low co-occurrence. Here, I examined threshold values ranging from 1% (the most extreme cases of partners and excluders), to 50% (77 species). Using all threshold values, I found that in at least 80% of species the mean competition index with partners is greater than with excluders. I also compared the mean metabolic competition index of partners to the mean metabolic competition index of excluders in different phylogenetic distance bins, using each threshold definition. I found that at any

threshold less extreme than 25%, species have significantly greater metabolic competition with partners than with excluders in any phylogenetic bin ($p < 0.05$ in all bins, one-tailed Mann-Whitney U test). Using thresholds of 15% or 20%, species have significantly greater metabolic competition with partners than with excluders in all bins but that of the lowest phylogenetic distance. For more extreme threshold values, metabolic competition still tends to be greater with partners than with excluders, but since fewer pairs of species are placed in each bin, significance could not be well established.

6.2. Supporting tables

Table 6.1: Oral strains analyzed, with genome sequence and characteristic colonization time

Species	Strain in growth assays	Genome used*	Colonization
Aa	Aggregatibacter actinomycetemcomitans JP2	Aggregatibacter actinomycetemcomitans D7S-1	Late
Ao	Actinomyces oris ATCC 43146	Actinomyces oris K20	Initial
Fn	Fusobacterium nucleatum ATCC 10953	Fusobacterium nucleatum polymorphum ATCC 10953	Middle
Pg	Porphyromonas gingivalis ATCC 33277	Porphyromonas gingivalis ATCC 33277	Early
Sg	Streptococcus gordonii DL1	Streptococcus gordonii str. Challis substr. CH1	Initial
So	Streptococcus oralis 34	Streptococcus oralis SK23, ATCC 35037	Initial
Va	Veillonella sp. PK1910	Veillonella atypica ACS-134-V-Col7a	Early

* The genome sequence used for metabolic network reconstruction. When annotations were not available for the strain used in the original growth assay, the alternative strain listed was used.

Table 6.2: Metabolic competition index of human oral species

	Aa	Ao	Fn	Pg	Sg	So	Va
Aa		0.362	0.451	0.313	0.491	0.473	0.357
Ao	0.413		0.393	0.321	0.551	0.51	0.316
Fn	0.432	0.322		0.364	0.466	0.475	0.508
Pg	0.419	0.375	0.484		0.419	0.425	0.405
Sg	0.571	0.571	0.561	0.388		0.847	0.388
So	0.541	0.520	0.577	0.383	0.847		0.408
Va	0.435	0.335	0.652	0.359	0.413	0.435	

Table 6.3: Metabolic complementarity index of human oral species

	Aa	Ao	Fn	Pg	Sg	So	Va
Aa		0.161	0.143	0.089	0.089	0.107	0.161
Ao	0.224		0.143	0.143	0.163	0.184	0.204
Fn	0.237	0.169		0.119	0.102	0.119	0.136
Pg	0.310	0.238	0.31		0.19	0.214	0.310
Sg	0.184	0.163	0.122	0.122		0.02	0.143
So	0.163	0.184	0.122	0.122	0.020		0.143
Va	0.196	0.152	0.174	0.152	0.174	0.174	

Table 6.4: Oral growth assay trios.

Target species	Partner 1	Partner 2	Fold change 1*	Fold change 2*	Growth assessment**
Aa	Ao	Fn		4	
	Ao	Pg		4	
	Ao	So		0	
	Ao	Va		2	
	Fn	Pg	4	4	Pg > Fn
	Fn	Sg	4		
	Fn	So	4	0	
	Fn	Va	4	2	Fn > Va
	Pg	Sg	4		
	Pg	So	4	0	
	Pg	Va	4	2	Pg > Va
	Sg	So		0	
	Sg	Va		2	
	So	Va	0	2	Va > So
Ao	Aa	Fn		5	
	Aa	Pg		2	
	Aa	Sg		0	
	Aa	So		8	
	Aa	Va		7	
	Fn	Pg	5	2	Fn > Pg

	Fn	Sg	5	0	Fn > Sg
	Fn	So	5	8	So > Fn
	Fn	Va	5	7	Va > Fn
	Pg	Sg	2	0	Pg > Sg
	Pg	So	2	8	So > Pg
	Pg	Va	2	7	Va > Pg
	Sg	So	0	8	So > Sg
	Sg	Va	0	7	Va > Sg
	So	Va	8	7	
Fn	Aa	Ao	6	9	Ao > Aa
	Aa	Pg	6	9	Pg > Aa
	Aa	Sg	6		
	Aa	So	6	0	Aa > So
	Aa	Va	6	3	Aa > Va
	Ao	Pg	9	9	
	Ao	Sg	9		
	Ao	So	9	0	Ao > So
	Ao	Va	9	3	Ao > Va
	Pg	Sg	9		
	Pg	So	9	0	Pg > So
	Pg	Va	9	3	Pg > Va
	Sg	So		0	
	Sg	Va		3	
	So	Va	0	3	Va > So
Pg	Aa	Ao	4	12	Ao > Aa
	Aa	Fn	4	3	Aa > Fn
	Aa	Sg	4	3	Aa > Sg
	Aa	So	4	0	Aa > So
	Aa	Va	4	5	Va > Aa
	Ao	Fn	12	3	Ao > Fn
	Ao	Sg	12	3	Ao > Sg
	Ao	So	12	0	Ao > So
	Ao	Va	12	5	Ao > Va
	Fn	Sg	3	3	Fn > Sg
	Fn	So	3	0	Fn > So
	Fn	Va	3	5	Va > Fn
	Sg	So	3	0	Sg > So
	Sg	Va	3	5	Va > Sg
	So	Va	0	5	Va > So
Sg	Aa	Ao		1.5	
	Aa	Pg		2	
	Aa	So		1.5	
	Ao	Fn	1.5		
	Ao	Pg	1.5	2	
	Ao	So	1.5	1.5	
	Ao	Va	1.5		
	Fn	Pg		2	
	Fn	So		1.5	
	Pg	So	2	1.5	
	Pg	Va	2		

	So	Va	1.5		
So	Aa	Ao	0	2	
	Aa	Fn	0	9	
	Aa	Pg	0	0	
	Aa	Sg	0	0	
	Aa	Va	0	8	
	Ao	Fn	2	9	Fn > Ao
	Ao	Pg	2	0	Ao > Pg
	Ao	Sg	2	0	Ao > Sg
	Ao	Va	2	8	Va > Ao
	Fn	Pg	9	0	Fn > Pg
	Fn	Sg	9	0	Fn > Sg
	Fn	Va	9	8	Fn > Va
	Pg	Sg	0	0	
	Pg	Va	0	8	Va > Pg
	Sg	Va	0	8	Va > Sg
Va	Aa	Ao	3	7	Ao > Aa
	Aa	Fn	3	3	
	Aa	Pg	3	5	Pg > Aa
	Aa	Sg	3		
	Aa	So	3	8	So > Aa
	Ao	Fn	7	3	Ao > Fn
	Ao	Pg	7	5	Ao > Pg
	Ao	Sg	7		
	Ao	So	7	8	So > Ao
	Fn	Pg	3	5	Pg > Fn
	Fn	Sg	3		
	Fn	So	3	8	So > Fn
	Pg	Sg	5		
	Pg	So	5	8	So > Pg
	Sg	So		8	

* The fold change of the target species between 18h and 48h, from figure 3 of Kolenbrander 2011.
** A more stringent assessment based on Palmer *et al.* 2001, Chalmers *et al.* 2008, Periasamy *et al.* 2009, Periasamy & Kolenbrander 2009 a & b, and Periasamy & Kolenbrander 2010.

Table 6.5: Intestinal strains included in the analysis

Genome in abundance mapping*	Genome used for our analysis**
Bacteroides uniformis	Bacteroides uniformis ATCC 8492
Alistipes putredinis	Alistipes putredinis DSM 17216
Parabacteroides merdae	Parabacteroides merdae ATCC 43184
Dorea longicatena	Dorea longicatena DSM 13814
Bacteroides caccae	Bacteroides caccae ATCC 43185
Ruminococcus bromii L2-63	Ruminococcus bromii L2 63
Bacteroides thetaiotaomicron VPI-5482	Bacteroides thetaiotaomicron VPI 5482
Clostridium sp SS2-1	Clostridium sp SS2 1
Eubacterium hallii	Eubacterium hallii DSM 3353
Ruminococcus lactaris	Ruminococcus lactaris ATCC 29176
Ruminococcus sp SR1 5	Ruminococcus sp SR1 5
Bifidobacterium adolescentis	Bifidobacterium adolescentis L2 32
Ruminococcus torques L2-14	Ruminococcus torques L2 14
Akkermansia muciniphila ATCC BAA-835	Akkermansia muciniphila ATCC BAA 835
unknown sp SS3 4	Clostridiales sp SS3 4
Butyrivibrio crossotus	Butyrivibrio crossotus DSM 2876
Dorea formicigenerans	Dorea formicigenerans ATCC 27755
Faecalibacterium prausnitzii SL3 3	Faecalibacterium prausnitzii SL3 3
Bacteroides cellulosilyticus	Bacteroides cellulosilyticus DSM 14838
Bacteroides stercoris	Bacteroides stercoris ATCC 43183
Collinsella aerofaciens	Collinsella aerofaciens ATCC 25986
Clostridium sp L2-50	Clostridium sp L2 50
Bacteroides eggerthii	Bacteroides eggerthii DSM 20697

Bacteroides vulgatus ATCC 8482	Bacteroides vulgatus ATCC 8482
Bacteroides sp. 2 1 7	Bacteroides sp 2 1 7
Roseburia intestinalis M50 1	Roseburia intestinalis M50 1
Bifidobacterium longum subsp. infantis CCUG 52486	Bifidobacterium longum longum CCUG 52486
Eubacterium siraeum 70 3	Eubacterium siraeum 70 3
Bacteroides ovatus	Bacteroides ovatus ATCC 8483
Prevotella copri	Prevotella copri CB7 DSM 18205
Eubacterium ventriosum	Eubacterium ventriosum ATCC 27560
Bacteroides coprocola	Bacteroides coprocola M16 DSM 17136
Bacteroides sp. 9 1 42FAA	Bacteroides sp 9 1 42FAA
Parabacteroides distasonis ATCC 8503	Parabacteroides distasonis ATCC 8503
Bacteriodes xylanisolvens XB1A	Bacteroides xylanisolvens XB1A
Bacteroides sp. 2 2 4	Bacteroides sp 2 2 4
Bacteroides sp. 4 3 47FAA	Bacteroides sp 4 3 47FAA
Eubacterium biforme	Eubacterium biforme DSM 3989
Bacteroides plebeius	Bacteroides plebeius M12 DSM 17135
Eubacterium rectale M104 1	Eubacterium rectale M104 1
Coprococcus eutactus	Coprococcus eutactus ATCC 27759
Coprococcus comes SL7 1	Coprococcus comes ATCC 27758
Bacteroides sp. D1	Bacteroides sp D1
Bacteroides finegoldii	Bacteroides finegoldii DSM 17565
Bacteroides intestinalis	Bacteroides intestinalis 341 DSM 17393
Ruminococcus obeum A2-162	Ruminococcus obeum A2 162
Parabacteroides johnsonii	Parabacteroides johnsonii DSM 18315
Bacteroides sp. D4	Bacteroides dorei 5 1 36 D4
Streptococcus thermophilus LMD-9	Streptococcus thermophilus LMD 9

Bacteroides pectinophilus	Bacteroides pectinophilus ATCC 43243
Clostridium leptum	Clostridium leptum DSM 753
Bacteroides dorei	Bacteroides dorei DSM 17855
Catenibacterium mitsuokai	Catenibacterium mitsuokai DSM 15897
Bacteroides coprophilus	Bacteroides coprophilus DSM 18228
Bifidobacterium bifidum NCIMB 41171	Bifidobacterium bifidum NCIMB 41171
Clostridium bolteae	Clostridium bolteae ATCC BAA 613
Ruminococcus gnavus	Ruminococcus gnavus ATCC 29149
Desulfovibrio piger ATCC29098	Desulfovibrio piger ATCC 29098
Bacteroides fragilis 3 1 12	Bacteroides fragilis 3 1 12
Bifidobacterium pseudocatenulatum	Bifidobacterium pseudocatenulatum DSM 20438
Clostridium sp M62 1	Clostridium sp M62 1
Escherichia coli O157:H7 str. EC4115	Escherichia coli O157 H7 EC4115
Holdemania filiformis	Holdemania filiformis VPI J1 31B 1 DSM 12042
Bifidobacterium catenulatum	Bifidobacterium catenulatum DSM 16992
Clostridium bartlettii	Clostridium bartlettii DSM 16795
Bacteroides capillosus	Pseudoflavonifractor capillosus ATCC 29799
Subdoligranulum variabile	Subdoligranulum variabile DSM 15176
Clostridium symbiosum	Clostridium symbiosum WAL 14163
Mitsuokella multacida	Mitsuokella multacida DSM 20544
Clostridium nexile	Clostridium nexile DSM 1787
Methanobrevibacter smithii DSM2375	Methanobrevibacter smithii DSM 2375
Anaerotruncus colihominis	Anaerotruncus colihominis DSM 17241
Gordonibacter pamelaeae gen nov sp Nov	Gordonibacter pamelaeae 7 10 1 bT DSM 19378
Megamonas hypermegale ART12 1	Megamonas hypermegale ART12 1
Blautia hansenii	Blautia hansenii VPI C7 24 DSM 20583

Clostridium spiroforme	Clostridium spiroforme DSM 1552
Bifidobacterium animalis subsp. lactis AD011	Bifidobacterium animalis lactis AD011
Clostridium asparagiforme	Clostridium asparagiforme DSM 15981
Clostridium scindens	Clostridium scindens ATCC 35704
Bifidobacterium dentium	Bifidobacterium dentium ATCC 27678
Enterococcus faecalis TX0104	Enterococcus faecalis TX0104
Blautia hydrogenotrophica	Blautia hydrogenotrophica DSM 10507
Bifidobacterium angulatum	Bifidobacterium angulatum DSM 20098
Streptococcus gordonii str. Challis substr. CH1	Streptococcus gordonii Challis CH1
Clostridium methylpentosum	Clostridium methylpentosum R2 DSM 5476
Mollicutes bacterium D7	Coprobacillus sp D7
Eubacterium dolichum	Eubacterium dolichum DSM 3991
Streptococcus pneumoniae Hungary19A-6	Streptococcus pneumoniae Hungary19A 6
Bifidobacterium breve	Bifidobacterium breve DSM 20213
Lactobacillus salivarius ATCC 11741	Lactobacillus salivarius HO66 ATCC 11741
Lactobacillus acidophilus NCFM	Lactobacillus acidophilus NCFM
Klebsiella pneumoniae 342	Klebsiella pneumoniae 342
Lactobacillus johnsonii NCC 533	Lactobacillus johnsonii NCC 533
Collinsella stercoris	Collinsella stercoris DSM 13279
Bryantella formatexigens	Bryantella formatexigens I 52 DSM 14469
Anaerostipes caccae	Anaerostipes caccae DSM 14662
Streptococcus sanguinis SK36	Streptococcus sanguinis SK36
Clostridium ramosum	Clostridium ramosum VPI 0427 DSM 1402
Streptococcus infantarius	Streptococcus infantarius infantarius ATCC BAA 102
Pediococcus pentosaceus ATCC 25745	Pediococcus pentosaceus ATCC 25745
Streptococcus mutans UA159	Streptococcus mutans UA159

Escherichia fergusonii ATCC 35469	Escherichia fergusonii UMN026 ATCC 35469
Haemophilus influenzae 86-028NP	Haemophilus influenzae NTHi 86 028NP
Collinsella intestinalis	Collinsella intestinalis DSM 13280
Actinomyces odontolyticus	Actinomyces odontolyticus ATCC 17982
Helicobacter pullorum MIT 98-5489	Helicobacter pullorum MIT 98 5489
Butyrivibrio fibrisolvens 16 4	Butyrivibrio fibrisolvens 16 4
Lactobacillus delbrueckii subsp. bulgaricus ATCC 11842	Lactobacillus delbrueckii bulgaricus ATCC 11842
Enterobacter cancerogenus	Enterobacter cancerogenus ATCC 35316
Lactobacillus sakei subsp. sakei 23K	Lactobacillus sakei sakei 23K
Lactobacillus gasseri ATCC 33323	Lactobacillus gasseri ATCC 33323
Fusobacterium nucleatum subsp. nucleatum ATCC 25586	Fusobacterium nucleatum nucleatum ATCC 25586
Lactococcus lactis subsp. cremoris MG1363	Lactococcus lactis cremoris MG1363
Candidatus Sulcia muelleri GWSS	Candidatus Sulcia muelleri GWSS
Clostridium difficile 630	Clostridium difficile 630
Lactobacillus fermentum IFO 3956	Lactobacillus fermentum IFO 3956
Clostridium hylemonae	Clostridium hylemonae DSM 15053
Methanosphaera stadtmanae DSM 3091	Methanosphaera stadtmanae DSM 3091
Campylobacter hominis ATCC BAA-381	Campylobacter hominis ATCC BAA 381
Porphyromonas gingivalis ATCC 33277	Porphyromonas gingivalis ATCC 33277
Lactobacillus casei BL23	Lactobacillus casei casei BL23
Clostridium phytofermentans ISDg	Clostridium phytofermentans ISDg
Clostridium sp. 7 2 43FAA	Clostridium sp 7 2 43FAA
Enterococcus sp 7L76	Enterococcus sp 7L76
Enterococcus faecalis TX1332	Enterococcus faecalis TX1322
Clostridium perfringens ATCC 13124	Clostridium perfringens ATCC 13124
Leuconostoc mesenteroides ATCC 8293	Leuconostoc mesenteroides mesenteroides ATCC 8293

Citrobacter sp. 30 2	Citrobacter sp 30 2
Lactobacillus ultunensis DSM 16047	Lactobacillus ultunensis DSM 16047
Citrobacter koseri ATCC BAA-895	Citrobacter koseri ATCC BAA 895
Desulfovibrio vulgaris str. 'Miyazaki F'	Desulfovibrio vulgaris Miyazaki F
Proteus mirabilis HI4320	Proteus mirabilis HI4320
Lactobacillus helveticus DPC 4571	Lactobacillus helveticus DPC 4571
Citrobacter sp	n/a***
Enterobacter sp. 638	Enterobacter sp 638
Anaerofustis stercorihominis	Anaerofustis stercorihominis DSM 17244
Finegoldia magna ATCC 29328	Finegoldia magna ATCC 29328
Campylobacter concisus 13826	Campylobacter concisus 13826
Lactobacillus reuteri SD2112	Lactobacillus reuteri SD2112 ATCC 55730
Streptococcus suis 05ZYH33	Streptococcus suis 05ZYH33
Tropheryma whipplei str. Twist	Tropheryma whipplei Twist
Pseudomonas aeruginosa LESB58	Pseudomonas aeruginosa LESB58
Salmonella enterica serovar Heidelberg str. SL476	Salmonella enterica sv Heidelberg SL476 CVM30485
Bifidobacterium gallicum	Bifidobacterium gallicum DSM 20093
Lactobacillus paracasei subsp. paracasei ATCC 25302	Lactobacillus paracasei ATCC 25302
Pasteurella multocida subsp. multocida str. Pm70	Pasteurella multocida multocida Pm70
Streptococcus pyogenes MGAS10750	Streptococcus pyogenes M4 MGAS10750
Enterobacter sakazakii ATCC BAA-894	Cronobacter sakazakii ATCC BAA 894
Anaerococcus hydrogenalis	Anaerococcus hydrogenalis DSM 7454
Thermoanaerobacter sp. X514	Thermoanaerobacter sp X514
Haemophilus parasuis SH0165	Haemophilus parasuis SH0165
Lactobacillus hilgardii ATCC 8290	Lactobacillus hilgardii ATCC 8290
Actinobacillus pleuropneumoniae serovar 7	Actinobacillus pleuropneumoniae sv 7 AP76

str. AP76	
Proteus penneri	Proteus penneri ATCC 35198
Staphylococcus saprophyticus ATCC 15305	Staphylococcus saprophyticus saprophyticus ATCC 15305

* The strain used to map shotgun reads and to determine abundance across 124 metagenome samples, as listed in the supplementary information of Qin *et al.* 2010.

** The genome sequence used to reconstruct genome-scale metabolic networks and to predict metabolic interactions. When annotations were not available for the strain used in the original mapping, the alternative strain listed was used.

*** Citrobacter sp. did not have an unambiguous species listed, and was excluded from our analysis.

Table 6.6: Ecological traits of human intestinal species

Species	Oxygen Req.*	Habitat*	Temperature*	Pathogenicity**
Actinobacillus pleuropneumoniae sv 7 AP76	Facultative	Host-associated	Mesophilic	Porcine
Actinomyces odontolyticus ATCC 17982	Facultative	Host-associated	Mesophilic	Human
Akkermansia muciniphila ATCC BAA 835	Anaerobic	Host-associated	Mesophilic	
Alistipes putredinis DSM 17216	Anaerobic	Multiple	Mesophilic	Human, Sheep
Anaerococcus hydrogenalis DSM 7454	Anaerobic	Host-associated	Mesophilic	
Anaerofustis stercorihominis DSM 17244	Anaerobic	Host-associated	Mesophilic	No

Anaerostipes caccae DSM 14662	Anaerobic	Host-associated	Mesophilic	No
Anaerotruncus colihominis DSM 17241	Anaerobic	Host-associated	Mesophilic	No
Bacteroides caccae ATCC 43185	Anaerobic	Host-associated	Mesophilic	Human
Bacteroides cellulosilyticus DSM 14838	Anaerobic	Host-associated	Mesophilic	
Bacteroides coprocola M16 DSM 17136	Anaerobic	Host-associated	Mesophilic	No
Bacteroides coprophilus DSM 18228	Anaerobic	Host-associated	Mesophilic	No
Bacteroides dorei 5 1 36 D4	Anaerobic	Host-associated	Mesophilic	
Bacteroides dorei DSM 17855	Anaerobic	Host-associated	Mesophilic	No
Bacteroides eggerthii DSM 20697	Anaerobic	Host-associated	Mesophilic	No
Bacteroides finegoldii DSM 17565	Anaerobic	Host-associated	Mesophilic	No
Bacteroides fragilis 3 1 12	Anaerobic	Host-associated	Mesophilic	
Bacteroides intestinalis 341 DSM 17393	Anaerobic	Host-associated	Mesophilic	No
Bacteroides ovatus ATCC 8483	Anaerobic	Host-associated	Mesophilic	No

Bacteroides pectinophilus ATCC 43243	Anaerobic	Host-associated	Mesophilic	No
Bacteroides plebeius M12 DSM 17135	Anaerobic	Host-associated	Mesophilic	No
Bacteroides sp 2 1 7	Anaerobic		Mesophilic	
Bacteroides sp 2 2 4	Anaerobic	Host-associated	Mesophilic	
Bacteroides sp 4 3 47FAA	Anaerobic	Host-associated	Mesophilic	
Bacteroides sp 9 1 42FAA	Anaerobic	Host-associated	Mesophilic	
Bacteroides sp D1	Anaerobic	Host-associated	Mesophilic	
Bacteroides stercoris ATCC 43183	Anaerobic	Host-associated	Mesophilic	No
Bacteroides thetaiotaomicron VPI 5482	Anaerobic	Host-associated	Mesophilic	Mammal
Bacteroides uniformis ATCC 8492	Anaerobic	Host-associated	Mesophilic	No
Bacteroides vulgatus ATCC 8482	Anaerobic	Host-associated	Mesophilic	Mammal
Bacteroides xylanisolvens XB1A	Anaerobic		Mesophilic	
Bifidobacterium adolescentis L2 32	Anaerobic	Host-associated	Mesophilic	No

Bifidobacterium angulatum DSM 20098	Anaerobic	Host-associated	Mesophilic	No
Bifidobacterium animalis lactis AD011	Anaerobic	Multiple	Mesophilic	
Bifidobacterium bifidum NCIMB 41171	Anaerobic	Multiple	Mesophilic	No
Bifidobacterium breve DSM 20213	Anaerobic	Host-associated	Mesophilic	No
Bifidobacterium catenulatum DSM 16992	Anaerobic		Mesophilic	
Bifidobacterium dentium ATCC 27678	Anaerobic	Host-associated	Mesophilic	No
Bifidobacterium gallicum DSM 20093	Anaerobic	Host-associated	Mesophilic	
Bifidobacterium longum longum CCUG 52486	Anaerobic	Host-associated	Mesophilic	No
Bifidobacterium pseudocatenulatum DSM 20438	Anaerobic	Host-associated	Mesophilic	No
Blautia hansenii VPI C7 24 DSM 20583	Anaerobic	Host-associated	Mesophilic	No
Blautia hydrogenotrophica DSM 10507	Anaerobic		Mesophilic	
Bryantella formatexigens I 52 DSM 14469	Anaerobic	Host-associated	Mesophilic	No
Butyrivibrio crossotus DSM 2876	Anaerobic	Host-associated	Mesophilic	No

Butyrivibrio fibrisolvens 16 4	Anaerobic	Host-associated	Mesophilic	No
Campylobacter concisus 13826	Microaerophilic	Host-associated	Mesophilic	Human
Campylobacter hominis ATCC BAA 381	Anaerobic	Host-associated	Mesophilic	No
Candidatus Sulcia muelleri GWSS				
Catenibacterium mitsuokai DSM 15897	Anaerobic	Host-associated	Mesophilic	No
Citrobacter koseri ATCC BAA 895		Multiple	Mesophilic	Human, Animal
Citrobacter sp 30 2	Facultative	Host-associated	Mesophilic	
Clostridiales sp SS3 4	Anaerobic		Mesophilic	
Clostridium asparagiforme DSM 15981	Anaerobic	Host-associated	Psycrophilic	No
Clostridium bartlettii DSM 16795	Anaerobic	Host-associated	Mesophilic	No
Clostridium bolteae ATCC BAA 613	Anaerobic	Multiple	Mesophilic	
Clostridium difficile 630	Anaerobic	Multiple	Mesophilic	Human
Clostridium hylemonae DSM 15053	Anaerobic		Mesophilic	No

Clostridium leptum DSM 753	Anaerobic	Host-associated	Mesophilic	
Clostridium methylpentosum R2 DSM 5476	Anaerobic	Host-associated	Mesophilic	No
Clostridium nexile DSM 1787	Anaerobic	Multiple	Mesophilic	No
Clostridium perfringens ATCC 13124	Anaerobic	Multiple	Mesophilic	Human, Animal
Clostridium phytofermentans ISDg	Anaerobic	Terrestrial	Mesophilic	No
Clostridium ramosum VPI 0427 DSM 1402	Anaerobic	Host-associated	Mesophilic	Human
Clostridium scindens ATCC 35704	Anaerobic	Multiple	Mesophilic	
Clostridium sp 7 2 43FAA	Anaerobic		Mesophilic	
Clostridium sp L2 50	Anaerobic	Host-associated	Mesophilic	
Clostridium sp M62 1	Anaerobic	Host-associated	Mesophilic	
Clostridium sp SS2 1	Anaerobic	Host-associated	Mesophilic	
Clostridium spiroforme DSM 1552	Anaerobic	Host-associated	Mesophilic	Rabbit
Clostridium symbiosum WAL 14163	Anaerobic	Host-associated	Mesophilic	

Collinsella aerofaciens ATCC 25986	Anaerobic	Host-associated	Mesophilic	
Collinsella intestinalis DSM 13280	Anaerobic	Host-associated	Mesophilic	
Collinsella stercoris DSM 13279	Anaerobic	Host-associated	Mesophilic	No
Coprobacillus sp D7			Mesophilic	
Coprococcus comes ATCC 27758	Anaerobic	Host-associated	Mesophilic	
Coprococcus eutactus ATCC 27759	Anaerobic	Host-associated	Mesophilic	
Cronobacter sakazakii ATCC BAA 894	Anaerobic	Host-associated	Mesophilic	Human
Desulfovibrio piger ATCC 29098	Anaerobic	Multiple	Mesophilic	No
Desulfovibrio vulgaris Miyazaki F	Anaerobic	Multiple	Mesophilic	No
Dorea formicigenerans ATCC 27755	Anaerobic	Host-associated	Mesophilic	No
Dorea longicatena DSM 13814	Anaerobic	Host-associated	Mesophilic	
Enterobacter cancerogenus ATCC 35316	Facultative	Multiple	Mesophilic	Human
Enterobacter sp 638		Host-associated		

Enterococcus faecalis TX0104	Facultative	Multiple	Mesophilic	Human
Enterococcus faecalis TX1322***	Facultative	Multiple	Mesophilic	Human
Enterococcus sp 7L76				
Escherichia coli O157 H7 EC4115	Facultative	Multiple	Mesophilic	Human
Escherichia fergusonii UMN026 ATCC 35469	Facultative	Multiple	Mesophilic	Human, Animal
Eubacterium biforme DSM 3989	Anaerobic	Host-associated	Mesophilic	No
Eubacterium dolichum DSM 3991	Anaerobic	Host-associated	Mesophilic	No
Eubacterium hallii DSM 3353	Anaerobic	Host-associated	Mesophilic	No
Eubacterium rectale M104 1	Anaerobic		Mesophilic	
Eubacterium siraeum 70 3	Anaerobic	Host-associated	Mesophilic	No
Eubacterium ventriosum ATCC 27560	Anaerobic	Host-associated		No
Faecalibacterium prausnitzii SL3 3***	Anaerobic	Host-associated	Mesophilic	No
Finegoldia magna ATCC 29328	Anaerobic	Multiple	Mesophilic	Human

Fusobacterium nucleatum nucleatum ATCC 25586	Anaerobic	Host-associated	Mesophilic	Human, Animal
Gordonibacter pamelaeae 7 10 1 bT DSM 19378	Anaerobic			
Haemophilus influenzae NTHi 86 028NP	Facultative	Host-associated	Mesophilic	Human
Haemophilus parasuis SH0165	Facultative	Host-associated	Mesophilic	Porcine
Helicobacter pullorum MIT 98 5489	Microaerophilic	Multiple	Mesophilic	Human
Holdemania filiformis VPI J1 31B 1 DSM 12042	Anaerobic	Host-associated	Mesophilic	No
Klebsiella pneumoniae 342	Facultative	Host-associated	Mesophilic	No
Lactobacillus acidophilus NCFM	Facultative	Multiple	Mesophilic	No
Lactobacillus casei casei BL23	Facultative	Specialized	Mesophilic	No
Lactobacillus delbrueckii bulgaricus ATCC 11842***	Facultative	Multiple	Mesophilic	No
Lactobacillus fermentum IFO 3956	Facultative	Multiple	Mesophilic	
Lactobacillus gasseri ATCC 33323	Facultative	Host-associated	Mesophilic	No
Lactobacillus helveticus DPC 4571	Facultative	Multiple	Mesophilic	No

Lactobacillus hilgardii ATCC 8290	Microaerophilic		Mesophilic	
Lactobacillus johnsonii NCC 533	Facultative	Host-associated	Mesophilic	No
Lactobacillus paracasei ATCC 25302	Microaerophilic		Mesophilic	
Lactobacillus reuteri SD2112 ATCC 55730	Facultative	Multiple	Mesophilic	
Lactobacillus sakei sakei 23K	Facultative	Multiple	Mesophilic	
Lactobacillus salivarius HO66 ATCC 11741	Facultative	Multiple	Mesophilic	No
Lactobacillus ultunensis DSM 16047	Microaerophilic		Mesophilic	
Lactococcus lactis cremoris MG1363	Facultative	Multiple	Mesophilic	No
Leuconostoc mesenteroides mesenteroides ATCC 8293	Facultative	Multiple	Mesophilic	No
Megamonas hypermegale ART12 1	Anaerobic	Host-associated	Mesophilic	
Methanobrevibacter smithii DSM 2375	Anaerobic	Multiple	Mesophilic	No
Methanosphaera stadtmanae DSM 3091	Anaerobic	Host-associated	Mesophilic	
Mitsuokella multacida DSM 20544	Anaerobic	Multiple	Mesophilic	No

Parabacteroides distasonis ATCC 8503	Anaerobic	Host-associated	Mesophilic	Mammal
Parabacteroides johnsonii DSM 18315	Anaerobic	Host-associated	Mesophilic	
Parabacteroides merdae ATCC 43184	Anaerobic	Host-associated	Mesophilic	
Pasteurella multocida multocida Pm70	Facultative	Host-associated	Mesophilic	Human, Animal
Pediococcus pentosaceus ATCC 25745	Facultative	Multiple	Mesophilic	No
Porphyromonas gingivalis ATCC 33277	Anaerobic	Host-associated	Mesophilic	Human
Prevotella copri CB7 DSM 18205	Anaerobic	Host-associated	Mesophilic	
Proteus mirabilis HI4320	Anaerobic	Host-associated	Mesophilic	Human
Proteus penneri ATCC 35198	Anaerobic	Host-associated	Mesophilic	Human
Pseudoflavonifractor capillosus ATCC 29799	Anaerobic	Host-associated	Mesophilic	Human
Pseudomonas aeruginosa LESB58	Aerobic	Multiple	Mesophilic	Human
Roseburia intestinalis M50 1***	Anaerobic	Host-associated	Mesophilic	No
Ruminococcus bromii L2 63	Anaerobic		Mesophilic	

Ruminococcus gnavus ATCC 29149	Anaerobic	Host-associated	Mesophilic	
Ruminococcus lactaris ATCC 29176	Anaerobic	Host-associated	Mesophilic	
Ruminococcus obeum A2 162***	Anaerobic	Host-associated	Mesophilic	No
Ruminococcus sp SR1 5	Anaerobic	Host-associated	Mesophilic	
Ruminococcus torques L2 14***	Anaerobic	Host-associated	Mesophilic	No
Salmonella enterica sv Heidelberg SL476 CVM30485	Facultative	Multiple	Mesophilic	Human, Animal
Staphylococcus saprophyticus saprophyticus ATCC 15305	Aerobic	Host-associated	Mesophilic	Human
Streptococcus gordonii Challis CH1	Facultative	Host-associated	Mesophilic	Human
Streptococcus infantarius infantarius ATCC BAA 102	Facultative	Host-associated	Mesophilic	No
Streptococcus mutans UA159	Facultative	Host-associated	Mesophilic	Human
Streptococcus pneumoniae Hungary19A 6	Facultative	Multiple	Mesophilic	Human
Streptococcus pyogenes M4 MGAS10750	Facultative	Host-associated	Mesophilic	Human
Streptococcus sanguinis SK36	Facultative	Host-associated	Mesophilic	Human

Streptococcus suis 05ZYH33	Facultative	Multiple	Mesophilic	Human, Swine
Streptococcus thermophilus LMD 9	Facultative	Multiple	Mesophilic	No
Subdoligranulum variabile DSM 15176	Anaerobic	Host-associated	Mesophilic	
Thermoanaerobacter sp X514	Anaerobic		Thermophilic	No
Tropheryma whipplei Twist	Aerobic	Host-associated	Mesophilic	Human

* For analysis purposes, missing data in oxygen requirement, habitat, and temperature are considered unknown
** For analysis purposes, missing data in pathogenicity are considered non-pathogenic
*** Alternative strain chosen for ecological traits

Table 6.7: Correlation of co-occurrence with metabolic interaction indices

	# species*	Metabolic competition index		Metabolic complementarity index	
		ρ =	p <	ρ =	p <
All species, all samples	154	0.211	0.0001	-0.193	0.0001
Controlling for phylogeny*	143	0.133	0.0001	-0.128	0.0001

* Mantel test of Spearman partial correlation

Table 6.8: Correlation of co-occurrence with metabolic interaction indices within phyla

	# species*	Metabolic competition index		Metabolic complementarity index	
		ρ =	p <	ρ =	p <
Bacteroidetes:	29	0.467	0.0005	-0.325	0.003
Firmicutes:	84	0.233	0.0001	-0.229	0.0001
Proteobacteria:	21	0.328	0.018	-0.33	0.009

* Only phyla with at least 20 species were considered.

Table 6.9: Average Metabolic competition index among partners and excluders, binned by 16s relatedness

Bin	Center	Total # of species pairs	# of partners	# of excluders	Mean metabolic competition among partners	Mean metabolic competition among excluders	p <
1	0.498	506	96	76	0.247	0.221	0.048
2	0.590	608	213	124	0.543	0.453	$3.5x10^{-5}$
3	0.681	9744	2638	2131	0.446	0.444	$9.7x10^{-5}$
4	0.772	6930	1863	1601	0.479	0.464	$1.91x10^{-4}$
5	0.863	1820	158	732	0.675	0.555	$1.0x10^{-9}$
6	0.954	688	37	335	0.784	0.661	$1.0x10^{-9}$

Table 6.10: Correlation of co-occurrence and metabolic interaction indices, partitioned by host health, nationality, and enterotype

	# of samples	Metabolic competition Index		Metabolic complementarity index	
		ρ =	p <	ρ =	p <
Healthy & Lean	60	0.186	0.0001	-0.183	0.0001
Healthy & Obese	39	0.175	0.0001	-0.182	0.0001
IBD & Lean	22	0.232	0.0001	-0.192	0.0001
IBD & Obese	3	0.182	0.0001	-0.166	0.0001
Danish	85	0.173	0.0001	-0.186	0.0001
Spanish	39	0.237	0.0001	-0.192	0.0001
Enterotype 1*	63	0.173	0.0001	-0.179	0.0001
Enterotype 2*	9	0.124	0.0001	-0.187	0.0001
Enterotype 3*	13	0.146	0.0001	-0.175	0.0001

* Enterotypes are only defined for the Danish samples.

Table 6.11: HMP organisms and genomes used for this analysis

Genome in abundance mapping	Genome used for Reverse Ecology
Abiotrophia defectiva	Abiotrophia defectiva ATCC 49176
Acidaminococcus fermentans	Acidaminococcus fermentans DSM 20731
Acinetobacter baumannii	Acinetobacter baumannii 6013113
Acinetobacter johnsonii	Acinetobacter johnsonii SH046
Acinetobacter junii	Acinetobacter junii SH205

Acinetobacter lwoffii	Acinetobacter lwoffii SH145
Acinetobacter radioresistens	Acinetobacter radioresistens SK82
Actinobacillus minor	Actinobacillus minor NM305
Actinobacillus succinogenes	Actinobacillus succinogenes 130Z
Actinomyces odontolyticus	Actinomyces odontolyticus F0309
Actinomyces oris	Actinomyces oris K20
Actinomyces urogenitalis	Actinomyces urogenitalis DSM 15434
Actinomyces viscosus	Actinomyces viscosus C505
Aerococcus viridans	Aerococcus viridans ATCC 11563
Aggregatibacter actinomycetemcomitans	Aggregatibacter actinomycetemcomitans D7S-1
Aggregatibacter aphrophilus	Aggregatibacter aphrophilus NJ8700
Aggregatibacter segnis	Aggregatibacter segnis ATCC 33393
Akkermansia muciniphila	Akkermansia muciniphila ATCC BAA-835
Alistipes putredinis	Alistipes putredinis DSM 17216
Alistipes shahii	Alistipes shahii WAL 8301
Anaerococcus hydrogenalis	Anaerococcus hydrogenalis DSM 7454
Anaerococcus lactolyticus	Anaerococcus lactolyticus ATCC 51172
Anaerococcus tetradius	Anaerococcus tetradius ATCC 35098
Anaerococcus vaginalis	Anaerococcus vaginalis ATCC 51170
Anaerofustis stercorihominis	Anaerofustis stercorihominis DSM 17244
Anaerostipes caccae	Anaerostipes caccae DSM 14662
Anaerotruncus colihominis	Anaerotruncus colihominis DSM 17241
Atopobium parvulum	Atopobium parvulum DSM 20469
Atopobium rimae	Atopobium rimae ATCC 49626
Atopobium vaginae	Atopobium vaginae DSM 15829
Bacteroides caccae	Bacteroides caccae ATCC 43185
Bacteroides cellulosilyticus	Bacteroides cellulosilyticus DSM 14838
Bacteroides coprocola	Bacteroides coprocola DSM 17136
Bacteroides coprophilus	Bacteroides coprophilus DSM 18228
Bacteroides dorei	Bacteroides dorei DSM 17855
Bacteroides eggerthii	Bacteroides eggerthii 1 2 48FAA
Bacteroides finegoldii	Bacteroides finegoldii DSM 17565
Bacteroides fragilis	Bacteroides fragilis 3 1 12
Bacteroides helcogenes	Bacteroides helcogenes P 36-108
Bacteroides intestinalis	Bacteroides intestinalis DSM 17393
Bacteroides ovatus	Bacteroides ovatus SD CMC 3f

Bacteroides pectinophilus	Bacteroides pectinophilus ATCC 43243
Bacteroides plebeius	Bacteroides plebeius M12, DSM 17135
Bacteroides salanitronis	Bacteroides salanitronis DSM 18170
Bacteroides stercoris	Bacteroides stercoris ATCC 43183
Bacteroides thetaiotaomicron	Bacteroides thetaiotaomicron VPI-5482
Bacteroides uniformis	Bacteroides uniformis ATCC 8492
Bacteroides vulgatus	Bacteroides vulgatus PC510
Bacteroides xylanisolvens	Bacteroides xylanisolvens SD CC 1b
Bifidobacterium adolescentis	Bifidobacterium adolescentis L2-32
Bifidobacterium angulatum	Bifidobacterium angulatum DSM 20098
Bifidobacterium animalis	Bifidobacterium animalis subsp. lactis V9
Bifidobacterium bifidum	Bifidobacterium bifidum NCIMB 41171
Bifidobacterium breve	Bifidobacterium breve DSM 20213
Bifidobacterium catenulatum	Bifidobacterium catenulatum DSM 16992
Bifidobacterium dentium	Bifidobacterium dentium JCVIHMP022
Bifidobacterium longum	Bifidobacterium longum subsp. infantis ATCC 15697
Bifidobacterium pseudocatenulatum	Bifidobacterium pseudocatenulatum DSM 20438
Bilophila wadsworthia	Bilophila wadsworthia 3 1 6
Blautia hansenii	Blautia hansenii DSM 20583
Blautia hydrogenotrophica	Blautia hydrogenotrophica DSM 10507
Brevundimonas subvibrioides	Brevundimonas subvibrioides ATCC 15264
Brucella ceti	Brucella ceti M490 95 1
Marvinbryantia formatexigens	Bryantella formatexigens DSM 14469
Bulleidia extructa	Bulleidia extructa W1219
Burkholderia cenocepacia	Burkholderia cenocepacia HI2424
Butyrivibrio crossotus	Butyrivibrio crossotus DSM 2876
Campylobacter concisus	Campylobacter concisus 13826
Campylobacter curvus	Campylobacter curvus 525.92
Campylobacter gracilis	Campylobacter gracilis RM3268
Campylobacter hominis	Campylobacter hominis ATCC BAA-381
Campylobacter rectus	Campylobacter rectus RM3267
Campylobacter showae	Campylobacter showae RM3277
Capnocytophaga gingivalis	Capnocytophaga gingivalis ATCC 33624
Capnocytophaga ochracea	Capnocytophaga ochracea F0287
Capnocytophaga sputigena	Capnocytophaga sputigena Capno
Cardiobacterium hominis	Cardiobacterium hominis ATCC 15826

Catenibacterium mitsuokai	Catenibacterium mitsuokai DSM 15897
Catonella morbi	Catonella morbi ATCC 51271
Clostridium asparagiforme	Clostridium asparagiforme DSM 15981
Clostridium bartlettii	Clostridium bartlettii DSM 16795
Clostridium bolteae	Clostridium bolteae ATCC BAA-613
Clostridium botulinum	Clostridium botulinum H04402 065
Clostridium cf	Clostridium cf. saccharolyticum K10
Clostridium difficile	Clostridium difficile NAP08
Clostridium hathewayi	Clostridium hathewayi DSM 13479
Clostridium hiranonis	Clostridium hiranonis DSM 13275
Clostridium hylemonae	Clostridium hylemonae DSM 15053
Clostridium leptum	Clostridium leptum DSM 753
Clostridium methylpentosum	Clostridium methylpentosum DSM 5476
Clostridium nexile	Clostridium nexile DSM 1787
Clostridium perfringens	Clostridium perfringens B str. ATCC 3626
Clostridium saccharolyticum	Clostridium saccharolyticum WM1
Clostridium scindens	Clostridium scindens ATCC 35704
Clostridium symbiosum	Clostridium symbiosum WAL-14163
Collinsella aerofaciens	Collinsella aerofaciens ATCC 25986
Collinsella intestinalis	Collinsella intestinalis DSM 13280
Collinsella stercoris	Collinsella stercoris DSM 13279
Coprococcus catus	Coprococcus catus GD 7
Coprococcus comes	Coprococcus comes ATCC 27758
Coprococcus eutactus	Coprococcus eutactus ATCC 27759
Corynebacterium accolens	Corynebacterium accolens ATCC 49726
Corynebacterium amycolatum	Corynebacterium amycolatum SK46
Corynebacterium aurimucosum	Corynebacterium aurimucosum ATCC 700975
Corynebacterium genitalium	Corynebacterium genitalium ATCC 33030
Corynebacterium jeikeium	Corynebacterium jeikeium ATCC 43734
Corynebacterium kroppenstedtii	Corynebacterium kroppenstedtii DSM 44385
Corynebacterium matruchotii	Corynebacterium matruchotii ATCC 33806
Corynebacterium pseudogenitalium	Corynebacterium pseudogenitalium ATCC 33035
Corynebacterium striatum	Corynebacterium striatum ATCC 6940
Corynebacterium tuberculostearicum	Corynebacterium tuberculostearicum SK141
Cryptobacterium curtum	Cryptobacterium curtum DSM 15641
Deinococcus radiodurans	Deinococcus radiodurans R1

Delftia acidovorans	Delftia acidovorans SPH-1
Desulfovibrio desulfuricans	Desulfovibrio desulfuricans G20
Desulfovibrio piger	Desulfovibrio piger ATCC 29098
Dialister invisus	Dialister invisus DSM 15470
Dialister microaerophilus	Dialister microaerophilus UPII 345-E
Dorea formicigenerans	Dorea formicigenerans ATCC 27755
Dorea longicatena	Dorea longicatena DSM 13814
Eggerthella lenta	Eggerthella lenta VPI 0255, DSM 2243
Eikenella corrodens	Eikenella corrodens ATCC 23834
Enhydrobacter aerosaccus	Enhydrobacter aerosaccus SK60
Enterobacter cancerogenus	Enterobacter cancerogenus ATCC 35316
Enterobacter cloacae	Enterobacter cloacae subsp. cloacae NCTC 9394
Enterococcus faecalis	Enterococcus faecalis HH22
Enterococcus faecium	Enterococcus faecium DO
Enterococcus gallinarum	Enterococcus gallinarum EG2
Escherichia coli	Escherichia coli MS 124-1
Escherichia fergusonii	Escherichia fergusonii ATCC 35469
Eubacterium eligens	Eubacterium eligens ATCC 27750
Eubacterium hallii	Eubacterium hallii DSM 3353
Eubacterium limosum	Eubacterium limosum KIST612
Eubacterium rectale	Eubacterium rectale M104 1
Eubacterium saburreum	Eubacterium saburreum DSM 3986
Eubacterium saphenum	Eubacterium saphenum ATCC 49989
Eubacterium siraeum	Eubacterium siraeum DSM 15702
Eubacterium ventriosum	Eubacterium ventriosum ATCC 27560
Faecalibacterium cf	Faecalibacterium cf. prausnitzii KLE1255
Faecalibacterium prausnitzii	Faecalibacterium prausnitzii A2-165
Finegoldia magna	Finegoldia magna BVS033A4
Fusobacterium gonidiaformans	Fusobacterium gonidiaformans ATCC 25563
Fusobacterium mortiferum	Fusobacterium mortiferum ATCC 9817
Fusobacterium nucleatum	Fusobacterium nucleatum subsp. nucleatum ATCC 23726
Fusobacterium periodonticum	Fusobacterium periodonticum ATCC 33693
Fusobacterium ulcerans	Fusobacterium ulcerans ATCC 49185
Fusobacterium varium	Fusobacterium varium ATCC 27725
Gardnerella vaginalis	Gardnerella vaginalis ATCC 14019
Gemella haemolysans	Gemella haemolysans M341

Gemella moribillum	Gemella moribillum M424
Gordonibacter pamelaeae	Gordonibacter pamelaeae 7-10-1-b
Granulicatella adiacens	Granulicatella adiacens ATCC 49175
Granulicatella elegans	Granulicatella elegans ATCC 700633
Haemophilus influenzae	Haemophilus influenzae 3655
Haemophilus parainfluenzae	Haemophilus parainfluenzae ATCC 33392
Haemophilus parasuis	Haemophilus parasuis 29755
Helicobacter pylori	Helicobacter pylori 83
Herbaspirillum seropedicae	Herbaspirillum seropedicae SmR1
Histophilus somni	Histophilus somni 2336
Holdemania filiformis	Holdemania filiformis DSM 12042
Jonquetella anthropi	Jonquetella anthropi E3 33 E1
Kingella denitrificans	Kingella denitrificans ATCC 33394
Kingella oralis	Kingella oralis ATCC 51147
Klebsiella pneumoniae	Klebsiella pneumoniae subsp. rhinoscleromatis ATCC 13884
Klebsiella variicola	Klebsiella variicola At-22
Kocuria rhizophila	Kocuria rhizophila DC2201
Lactobacillus acidophilus	Lactobacillus acidophilus ATCC 4796
Lactobacillus brevis	Lactobacillus brevis subsp. gravesensis ATCC 27305
Lactobacillus casei	Lactobacillus casei LC2W
Lactobacillus coleohominis	Lactobacillus coleohominis 101-4-CHN
Lactobacillus crispatus	Lactobacillus crispatus 214-1
Lactobacillus delbrueckii	Lactobacillus delbrueckii subsp. bulgaricus PB2003 044-T3-4
Lactobacillus fermentum	Lactobacillus fermentum ATCC 14931
Lactobacillus gasseri	Lactobacillus gasseri 224-1
Lactobacillus helveticus	Lactobacillus helveticus DSM 20075
Lactobacillus iners	Lactobacillus iners LactinV 01V1-a
Lactobacillus jensenii	Lactobacillus jensenii 208-1
Lactobacillus oris	Lactobacillus oris PB013-T2-3
Lactobacillus paracasei	Lactobacillus paracasei subsp. paracasei 8700 2
Lactobacillus plantarum	Lactobacillus plantarum subsp. plantarum ATCC 14917
Lactobacillus reuteri	Lactobacillus reuteri SD2112
Lactobacillus rhamnosus	Lactobacillus rhamnosus LMS2-1
Lactobacillus ruminis	Lactobacillus ruminis ATCC 25644

Lactobacillus salivarius	Lactobacillus salivarius ACS-116-V-Col5a
Lactobacillus vaginalis	Lactobacillus vaginalis ATCC 49540
Lactococcus lactis	Lactococcus lactis subsp. cremoris MG1363
Lautropia mirabilis	Lautropia mirabilis ATCC 51599
Leptotrichia buccalis	Leptotrichia buccalis DSM 1135
Leptotrichia goodfellowii	Leptotrichia goodfellowii F0264
Leptotrichia hofstadii	Leptotrichia hofstadii F0254
Leuconostoc citreum	Leuconostoc citreum KM20
Leuconostoc gasicomitatum	Leuconostoc gasicomitatum LMG 18811
Leuconostoc mesenteroides	Leuconostoc mesenteroides subsp. cremoris ATCC 19254
Listeria monocytogenes	Listeria monocytogenes 08-5578
Mannheimia haemolytica	Mannheimia haemolytica serotype A2 str. OVINE
Megamonas hypermegale	Megamonas hypermegale ART12 1
Megasphaera micronuciformis	Megasphaera micronuciformis F0359
Methanobrevibacter smithii	Methanobrevibacter smithii DSM 2375
Methanosphaera stadtmanae	Methanosphaera stadtmanae DSM 3091
Methylobacterium radiotolerans	Methylobacterium radiotolerans JCM 2831
Micrococcus luteus	Micrococcus luteus SK58
Mitsuokella multacida	Mitsuokella multacida DSM 20544
Mobiluncus curtisii	Mobiluncus curtisii ATCC 43063
Moraxella catarrhalis	Moraxella catarrhalis RH4
Mycobacterium abscessus	Mycobacterium abscessus CIP 104536
Mycoplasma hominis	Mycoplasma hominis PG21, ATCC 23114
Neisseria cinerea	Neisseria cinerea ATCC 14685
Neisseria elongata	Neisseria elongata glycolytica ATCC 29315
Neisseria flavescens	Neisseria flavescens SK114
Neisseria gonorrhoeae	Neisseria gonorrhoeae NCCP11945
Neisseria lactamica	Neisseria lactamica ATCC 23970
Neisseria meningitidis	Neisseria meningitidis ATCC 13091
Neisseria mucosa	Neisseria mucosa ATCC 25996
Neisseria polysaccharea	Neisseria polysaccharea ATCC 43768
Neisseria sicca	Neisseria sicca ATCC 29256
Neisseria subflava	Neisseria subflava NJ9703
Odoribacter splanchnicus	Odoribacter splanchnicus 1651/6, DSM 20712
Olsenella uli	Olsenella uli VPI, DSM 7084
Oribacterium sinus	Oribacterium sinus F0268

Oxalobacter formigenes	Oxalobacter formigenes OXCC13
Pantoea ananatis	Pantoea ananatis LMG 20103
Parabacteroides distasonis	Parabacteroides distasonis ATCC 8503
Parabacteroides johnsonii	Parabacteroides johnsonii DSM 18315
Parabacteroides merdae	Parabacteroides merdae ATCC 43184
Paracoccus denitrificans	Paracoccus denitrificans PD1222
Parascardovia denticolens	Parascardovia denticolens DSM 10105
Parvimonas micros	Parvimonas micra ATCC 33270
Pasteurella dagmatis	Pasteurella dagmatis ATCC 43325
Pasteurella multocida	Pasteurella multocida subsp. multocida str. Pm70
Pediococcus acidilactici	Pediococcus acidilactici 7 4
Peptoniphilus duerdenii	Peptoniphilus duerdenii ATCC BAA-1640
Peptoniphilus harei	Peptoniphilus harei ACS-146-V-Sch2b
Peptoniphilus lacrimalis	Peptoniphilus lacrimalis 315-B
Peptostreptococcus anaerobius	Peptostreptococcus anaerobius 653-L
Peptostreptococcus stomatis	Peptostreptococcus stomatis DSM 17678
Porphyromonas asaccharolytica	Porphyromonas asaccharolytica PR426713P-I
Porphyromonas endodontalis	Porphyromonas endodontalis ATCC 35406
Porphyromonas gingivalis	Porphyromonas gingivalis ATCC 33277
Porphyromonas uenonis	Porphyromonas uenonis 60-3
Prevotella amnii	Prevotella amnii CRIS 21A-A
Prevotella bergensis	Prevotella bergensis DSM 17361
Prevotella bivia	Prevotella bivia JCVIHMP010
Prevotella buccae	Prevotella buccae D17
Prevotella buccalis	Prevotella buccalis ATCC 35310
Prevotella copri	Prevotella copri DSM 18205
Prevotella disiens	Prevotella disiens FB035-09AN
Prevotella marshii	Prevotella marshii DSM 16973
Prevotella melaninogenica	Prevotella melaninogenica D18
Prevotella multiformis	Prevotella multiformis DSM 16608
Prevotella oralis	Prevotella oralis ATCC 33269
Prevotella oris	Prevotella oris F0302
Prevotella salivae	Prevotella salivae DSM 15606
Prevotella tannerae	Prevotella tannerae ATCC 51259
Prevotella timonensis	Prevotella timonensis CRIS 5C-B1
Prevotella veroralis	Prevotella veroralis F0319

Prochlorococcus marinus	Prochlorococcus marinus str. MIT 9303
Propionibacterium acnes	Propionibacterium acnes J165
Propionibacterium freudenreichii	Propionibacterium freudenreichii subsp. shermanii CIRM-BIA1
Proteus mirabilis	Proteus mirabilis ATCC 29906
Pseudoflavonifractor capillosus	Pseudoflavonifractor capillosus ATCC 29799
Pseudomonas aeruginosa	Pseudomonas aeruginosa PAb1
Pseudomonas fluorescens	Pseudomonas fluorescens SBW25
Pseudomonas putida	Pseudomonas putida GB-1
Pseudoramibacter alactolyticus	Pseudoramibacter alactolyticus ATCC 23263
Psychrobacter arcticus	Psychrobacter arcticus 273-4
Psychrobacter cryohalolentis	Psychrobacter cryohalolentis K5
Pyramidobacter piscolens	Pyramidobacter piscolens W5455
Rhodobacter sphaeroides	Rhodobacter sphaeroides KD131
Rhodopseudomonas palustris	Rhodopseudomonas palustris TIE-1
Roseburia intestinalis	Roseburia intestinalis L1-82
Roseburia inulinivorans	Roseburia inulinivorans DSM 16841
Roseiflexus castenholzii	Roseiflexus castenholzii DSM 13941
Rothia dentocariosa	Rothia dentocariosa ATCC 17931
Rothia mucilaginosa	Rothia mucilaginosa ATCC 25296
Ruminococcus albus	Ruminococcus albus 8
Ruminococcus bromii	Ruminococcus bromii L2-63
Ruminococcus gnavus	Ruminococcus gnavus ATCC 29149
Ruminococcus lactaris	Ruminococcus lactaris ATCC 29176
Ruminococcus obeum	Ruminococcus obeum ATCC 29174
Ruminococcus torques	Ruminococcus torques L2-14
Selenomonas artemidis	Selenomonas artemidis F0399
Selenomonas flueggei	Selenomonas flueggei ATCC 43531
Selenomonas noxia	Selenomonas noxia ATCC 43541
Selenomonas sputigena	Selenomonas sputigena ATCC 35185
Serratia proteamaculans	Serratia proteamaculans 568
Shigella boydii	Shigella boydii Sb227
Shigella dysenteriae	Shigella dysenteriae 1617
Shigella sonnei	Shigella sonnei Ss046
Shuttleworthia satelles	Shuttleworthia satelles DSM 14600
Simonsiella muelleri	Simonsiella muelleri ATCC 29453

Slackia exigua	Slackia exigua ATCC 700122
Solobacterium moorei	Solobacterium moorei F0204
Sphingopyxis alaskensis	Sphingopyxis alaskensis RB2256
Staphylococcus aureus	Staphylococcus aureus subsp. aureus TCH130
Staphylococcus capitis	Staphylococcus capitis SK14
Staphylococcus epidermidis	Staphylococcus epidermidis M23864 W1
Staphylococcus haemolyticus	Staphylococcus haemolyticus JCSC1435
Staphylococcus hominis	Staphylococcus hominis SK119
Staphylococcus lugdunensis	Staphylococcus lugdunensis M23590
Staphylococcus pseudintermedius	Staphylococcus pseudintermedius HKU10-03
Staphylococcus saprophyticus	Staphylococcus saprophyticus subsp. saprophyticus ATCC 15305
Staphylococcus warneri	Staphylococcus warneri L37603
Streptobacillus moniliformis	Streptobacillus moniliformis DSM 12112
Streptococcus agalactiae	Streptococcus agalactiae COH1
Streptococcus anginosus	Streptococcus anginosus SK52
Streptococcus australis	Streptococcus australis ATCC 700641
Streptococcus bovis	Streptococcus bovis ATCC 700338
Streptococcus cristatus	Streptococcus cristatus ATCC 51100
Streptococcus dysgalactiae	Streptococcus dysgalactiae subsp. equisimilis GGS 124
Streptococcus gordonii	Streptococcus gordonii str. Challis substr. CH1
Streptococcus infantarius	Streptococcus infantarius subsp. infantarius ATCC BAA-102
Streptococcus infantis	Streptococcus infantis SK1076
Streptococcus mitis	Streptococcus mitis ATCC 6249
Streptococcus mutans	Streptococcus mutans NN2025
Streptococcus oralis	Streptococcus oralis SK23, ATCC 35037
Streptococcus parasanguinis	Streptococcus parasanguinis ATCC 903
Streptococcus peroris	Streptococcus peroris ATCC 700780
Streptococcus pneumoniae	Streptococcus pneumoniae TCH8431 19A
Streptococcus pyogenes	Streptococcus pyogenes ATCC 10782
Streptococcus salivarius	Streptococcus salivarius SK126
Streptococcus sanguinis	Streptococcus sanguinis VMC66
Streptococcus suis	Streptococcus suis 05ZYH33
Streptococcus thermophilus	Streptococcus thermophilus JIM 8232
Streptococcus vestibularis	Streptococcus vestibularis F0396

Subdoligranulum variabile	Subdoligranulum variabile DSM 15176
Sutterella wadsworthensis	Sutterella wadsworthensis 3 1 45B
Treponema denticola	Treponema denticola ATCC 35405
Treponema phagedenis	Treponema phagedenis F0421
Treponema vincentii	Treponema vincentii ATCC 35580
Ureaplasma parvum	Ureaplasma parvum serovar 3 str. ATCC 700970
Veillonella atypica	Veillonella atypica ACS-134-V-Col7a
Veillonella dispar	Veillonella dispar ATCC 17748
Veillonella parvula	Veillonella parvula ATCC 17745
Victivallis vadensis	Victivallis vadensis ATCC BAA-548
Xanthomonas campestris	Xanthomonas campestris pv. vesicatoria str. 85-10

Table 6.12: Correlation of co-occurrence and metabolic interaction indices, across and within five major body sites

Body Site	# of samples	# of species	Metabolic competition				Metabolic complementarity			
			Normal		16s Control		Normal		16s Control	
			ρ =	p <	ρ =	p <	ρ =	p <	ρ =	p <
Whole body	693	314	0.128	0.0001	0.156	0.0001	-0.074	0.0016	-0.094	0.0001
Airways	87	95	0.142	0.0002	0.104	0.0082	-0.119	0.0018	-0.079	0.0288
Gastrointestinal tract	139	204	0.103	0.0046	0.083	0.0155	-0.138	0.0001	-0.122	0.0004
Oral	382	206	0.072	0.0187	0.093	0.005	-0.035	0.1562	-0.053	0.0729
Skin	26	160	0.168	0.0001	0.161	0.0001	-0.078	0.0063	-0.063	0.0278
Urogenital tract	56	78	0.116	0.0012	0.104	0.0066	-0.129	0.0001	-0.12	0.0004

Table 6.13: Correlation of co-occurrence and metabolic interaction indices, using alternative co-occurrence metrics

	Metabolic competition index		Metabolic complementarity index	
	ρ =	p <	ρ =	p <
Bray-Curtis	0.211	0.0001	-0.193	0.0001
Jaccard	0.211	0.0001	-0.193	0.0001
Morisita-Horn	0.147	0.0018	-0.115	0.0033
Cosine similarity	0.149	0.0008	-0.111	0.0037
Bray-Curtis*	0.134	0.0019	-0.131	0.0002
Jaccard*	0.134	0.0017	-0.131	0.0008
Morisita-Horn*	0.138	0.0021	-0.108	0.0046

* When using range-normalized abundances, the lowest observed relative abundance was scaled to zero, the highest to one, with all other scores scaled proportionately

Table 6.14: Hypergeometric test of interaction indices of partners and excluders within ecological traits

			Metabolic Competition Index			Metabolic Complementarity Index		
	M[1]	***K***[2]	***N***[3]	***X***[4]	***p* <**	***N***[3]	***X***[4]	***p* <**
Pathogens	154	39	127	31	0.635	133	35	0.163
Human pathogens	154	33	127	25	0.814	133	29	0.294
Anaerobes	148	106	122	88	0.290	128	92	0.322
Facultative anaerobes	148	34	122	30	0.098	128	29	0.494
Host-associated	136	99	113	84	0.126	119	87	0.297
Multiple Habitat	136	35	113	28	0.629	119	30	0.544

[1]: M, # of species with relevant ecological information

[2]: K, # of species labeled with this trait

[3]: *N*, # of species with greater (lower) metabolic competition (complementarity) among partners than excluders

[4]: *X*, # of species with this trait and greater (lower) metabolic competition (complementarity) among partners than excluders

6.3. Supporting figures

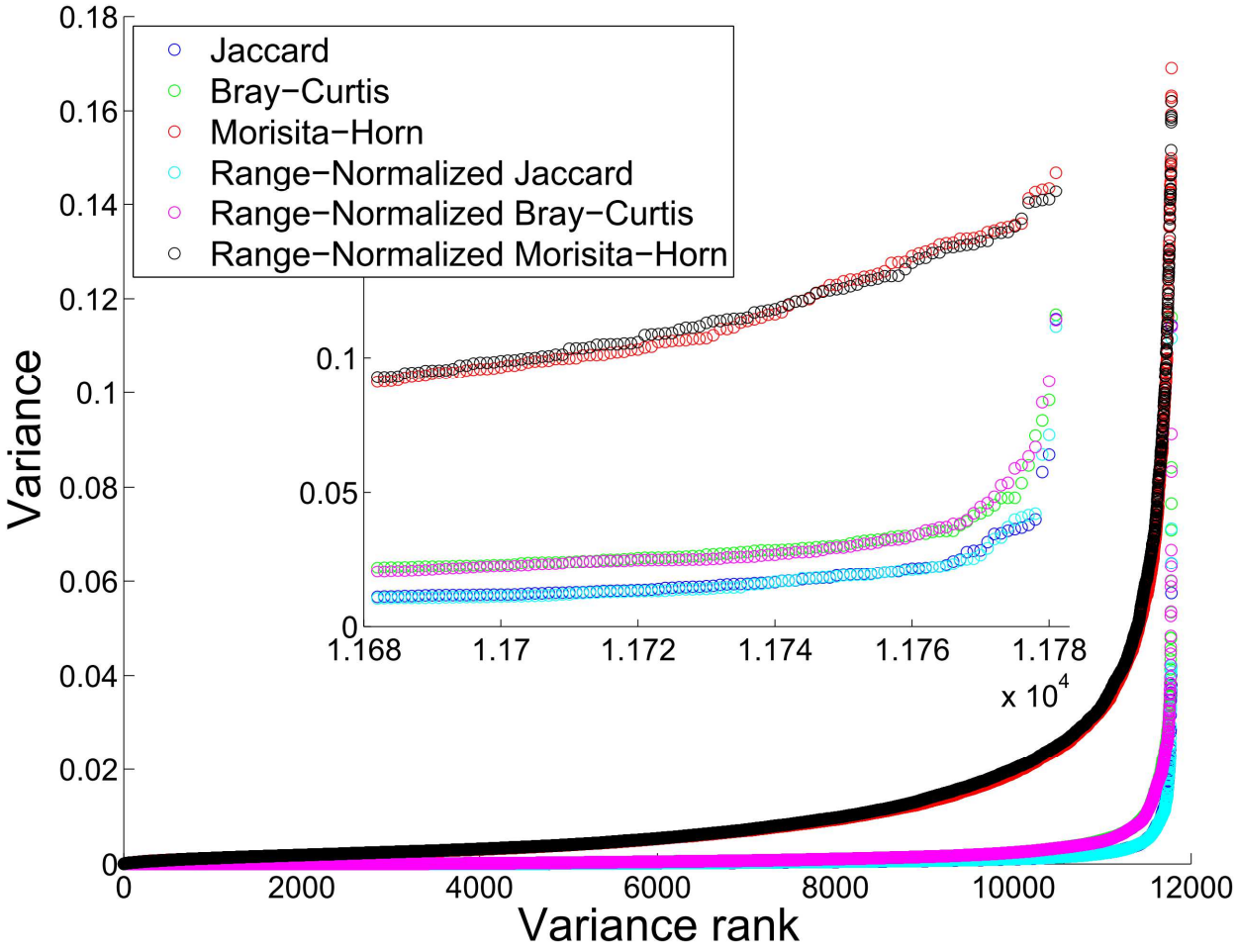


Figure 6.1: Robustness of co-occurrence metrics to under-sampling

Species abundances were subsampled 1000 times, from which co-occurrence of pairs was calculated. Variance across each species pair's 1000 co-occurrence values is plotted, with variance sorted from smallest to largest. Inset: Variance in Jaccard similarity index rises the slowest and has the lowest maximum of all co-occurrence metrics tested.

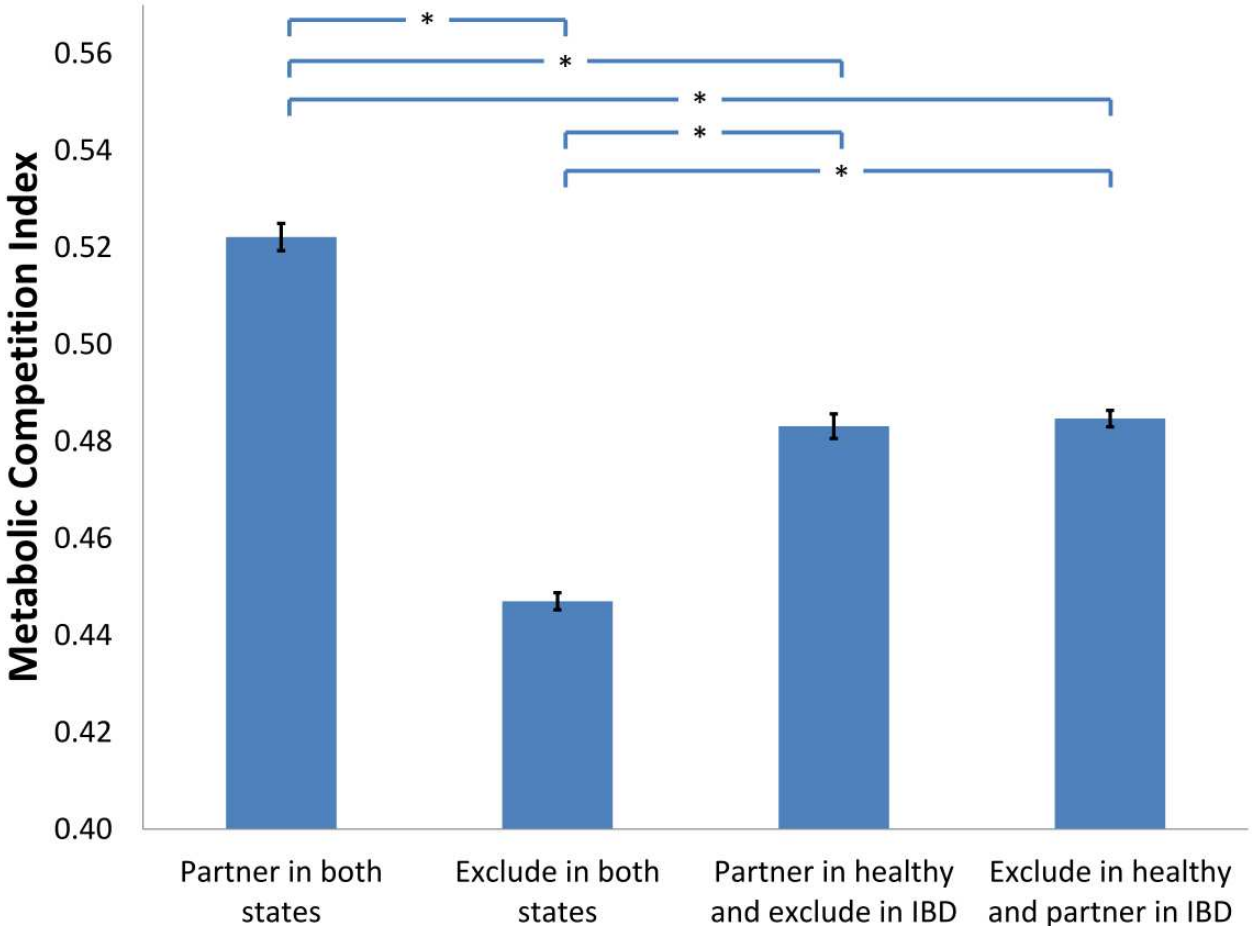


Figure 6.2: Metabolic competition index for consistent and inconsistent species co-occurrence

Bars represent the mean metabolic competition and standard error for species pairs that are found to be consistent partners or consistent excluders (*i.e.*, co-occur/exclude in both healthy and IBD samples), and for pairs that exhibit inconsistent co-occurrence patterns. Consistent partners have significantly different metabolic competition index from consistent excluders and from inconsistent pairs ($p < 0.05$; Wilcoxon rank-sum test).

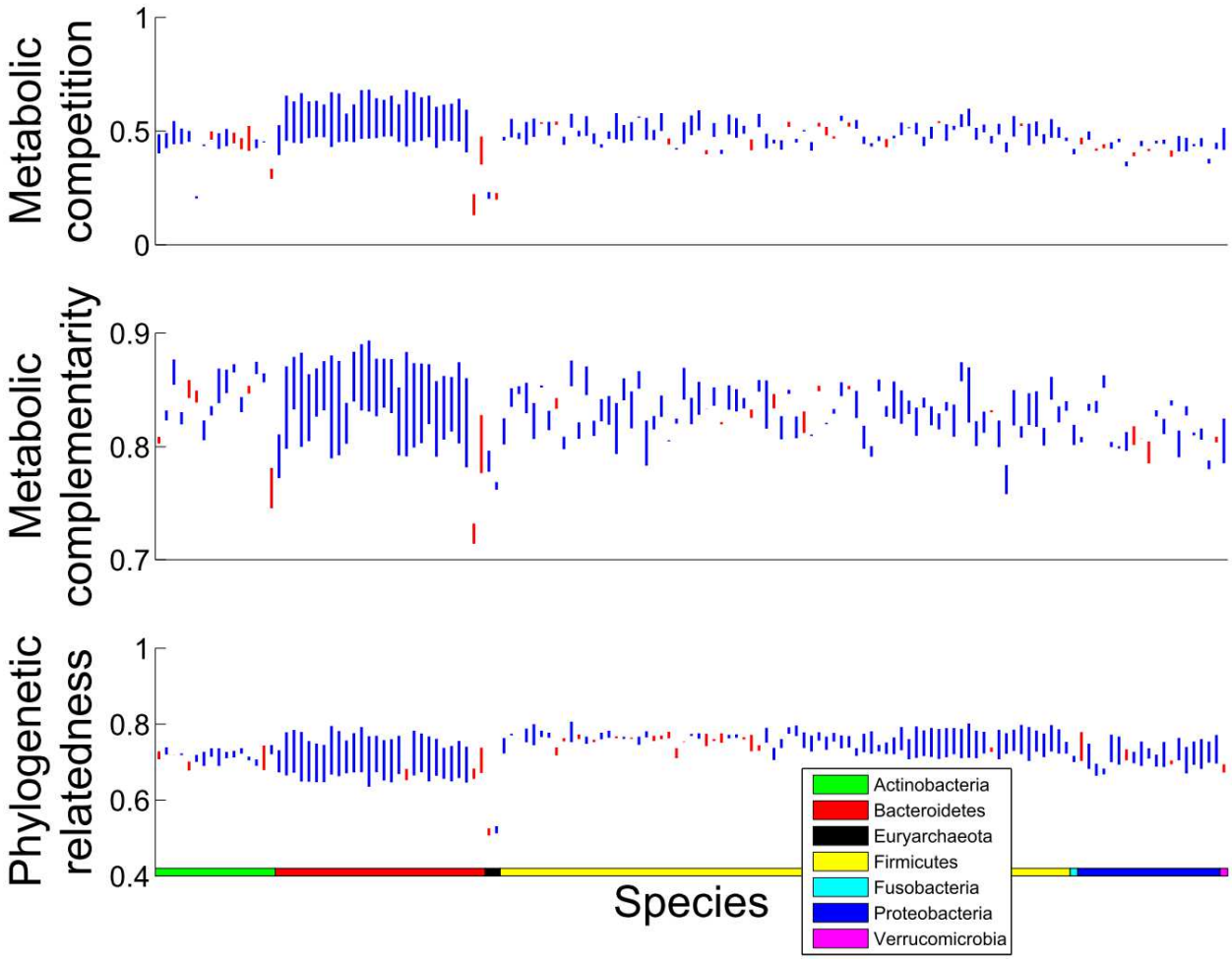


Figure 6.3: Comparison of competition, complementarity, and phylogeny in distinguishing partners vs. excluders in the human intestinal community.

Bars represent the range from the mean score among partners to the mean score among excluders. Blue bars indicate accurate separation, with partners having higher metabolic competition index, lower metabolic complementarity index, or higher phylogenetic relatedness than excluders.

7. Appendix B: Supporting Information for Chapter 3

7.1. Supporting text

7.1.1. Alternative definition of complementarity score

For the results reported in chapter 3, I defined the complementarity score of a pair of functions as the number of coincidences between OTUs of types *A* and type *B*; the same pair of OTUs may contribute to the complementarity score of a pair of functions many times if they are frequently found together in nature. This method weights more heavily functional complementarity which may be caused by dependency: organisms with intertwined lifestyles will be found together more frequently, or in more diverse environments than organisms which co-occur by chance. Nonetheless, there exists the possibility that generalists appear in many habitats, and therefore co-occur more than specialists, leading to a scenario in which this weighting method will heavily bias results to the strategies of metabolically versatile species.

To test whether this phenomenon influences our results, I tested an alternative approach in which each co-occurring pair of OTUs can only contribute once to the complementarity score of a function pair. Notably, this distribution of *p*-values is more heavily skewed towards 1.0 indicating that an appreciable fraction of significantly complemented function pairs is supported by frequently co-occurring microbes. Consequently, the number of highly significant complementarities (those with a minimal *p*-value of 10^{-4}) is greatly reduced (from 1,325 to 136). Therefore, in order to compare methods, I chose a less stringent threshold in order to achieve

roughly the same number of complements. Choosing $p < 0.05$ I arrive at a network with 1,373 edges connecting 261 nodes.

Once again the most common functional categories were transporters, two-component systems, and PTSs (with 73, 47, and 22 nodes, respectively). 56% (744) of the edges in the network reported in the main text also appear in this network. In spite of this apparently low overlap, 73% categorical enrichments are maintained (cf. Table 7.1, Table 3.1). Thus, while this scoring system generates a significantly altered network of complementarity functions, the two networks share similar high-level properties; for example, both networks are enriched for edges between two-component systems, PTSs, and transporters (although the alternative network lacks enrichment for transporter–PTS complements specifically), indicating that either method maintains a strong signal for complementarity between functions at the microbe-environment interface.

7.2. Supporting Tables

Table 7.1: Edge types significantly enriched in the alternative network of significantly complemented function pairs

Category	Category	N	K	*p* <	FDR
Arginine and proline metabolism	Citrate cycle (TCA cycle)	9	45	0.00E+00	1.00E-06
Arginine and proline metabolism	Fatty acid metabolism	3	18	2.47E-04	1.17E-02
Arginine and proline metabolism	Histidine metabolism	5	18	0.00E+00	5.20E-05
Arginine and proline metabolism	Lipopolysaccharide biosynthesis	8	36	0.00E+00	1.00E-06
Arginine and proline metabolism	Lysine biosynthesis	8	63	2.00E-06	1.50E-04
Arginine and proline metabolism	Oxidative phosphorylation	8	108	1.38E-04	7.00E-03
Arginine and proline metabolism	Phenylalanine, tyrosine and tryptophan	6	45	1.40E-05	9.40E-04

	biosynthesis				
Arginine and proline metabolism	Porphyrin and chlorophyll metabolism	3	18	2.47E-04	1.17E-02
Arginine and proline metabolism	Riboflavin metabolism	3	9	1.20E-05	8.14E-04
Arginine and proline metabolism	Thiamine metabolism	3	9	1.20E-05	8.14E-04
Arginine and proline metabolism	Transporters	39	810	0.00E+00	1.00E-06
Arginine and proline metabolism	Vitamin B6 metabolism	3	9	1.20E-05	8.14E-04
Citrate cycle (TCA cycle)	Fatty acid metabolism	6	10	0.00E+00	0.00E+00
Citrate cycle (TCA cycle)	Lipopolysaccharide biosynthesis	6	20	0.00E+00	4.00E-06
Citrate cycle (TCA cycle)	Phenylalanine, tyrosine and tryptophan biosynthesis	3	25	9.25E-04	3.12E-02
Citrate cycle (TCA cycle)	Terpenoid backbone biosynthesis	4	30	1.72E-04	8.57E-03
Glycine, serine and threonine metabolism	Porphyrin and chlorophyll metabolism	3	6	1.00E-06	1.31E-04
Histidine metabolism	Oxidative phosphorylation	3	24	7.88E-04	2.93E-02
Lipopolysaccharide biosynthesis	Glycine, serine and threonine metabolism	6	12	0.00E+00	0.00E+00
Lipopolysaccharide biosynthesis	One carbon pool by folate	3	12	4.40E-05	2.66E-03
Lysine biosynthesis	Biotin metabolism	3	7	3.00E-06	2.67E-04
Lysine biosynthesis	Fatty acid metabolism	4	14	3.00E-06	2.54E-04
Lysine biosynthesis	Glycine, serine and threonine metabolism	4	21	2.80E-05	1.79E-03
Lysine biosynthesis	Histidine metabolism	4	14	3.00E-06	2.54E-04
Lysine biosynthesis	Lipopolysaccharide biosynthesis	11	28	0.00E+00	0.00E+00
Lysine biosynthesis	Phenylalanine, tyrosine and tryptophan biosynthesis	4	35	3.63E-04	1.56E-02
Lysine biosynthesis	Porphyrin and chlorophyll metabolism	6	14	0.00E+00	0.00E+00
Lysine biosynthesis	Purine metabolism	4	28	1.22E-04	6.32E-03
Lysine biosynthesis	Transporters	35	630	0.00E+00	0.00E+00
Phosphotransferase system (PTS)	Phosphotransferase system (PTS)	12	276	1.47E-03	4.80E-02

Terpenoid backbone biosynthesis	Glycosaminoglycan degradation	3	24	7.88E-04	2.93E-02
Transporters	Citrate cycle (TCA cycle)	20	450	7.10E-05	4.06E-03
Transporters	Glycine, serine and threonine metabolism	13	270	3.85E-04	1.63E-02
Transporters	Histidine metabolism	11	180	1.00E-04	5.63E-03
Transporters	Riboflavin metabolism	8	90	3.30E-05	2.06E-03
Transporters	Terpenoid backbone biosynthesis	30	540	0.00E+00	2.00E-06
Transporters	Thiamine metabolism	10	90	1.00E-06	6.40E-05
Transporters	Two-component system	183	6660	0.00E+00	0.00E+00
Two-component system	Phosphotransferase system (PTS)	53	1776	1.18E-04	6.25E-03
Two-component system	Two-component system	106	2701	0.00E+00	0.00E+00
Valine, leucine and isoleucine biosynthesis	Arginine and proline metabolism	15	27	0.00E+00	0.00E+00
Valine, leucine and isoleucine biosynthesis	Biotin metabolism	3	3	0.00E+00	0.00E+00
Valine, leucine and isoleucine biosynthesis	Carbon fixation pathways in prokaryotes	4	15	5.00E-06	3.54E-04
Valine, leucine and isoleucine biosynthesis	Cysteine and methionine metabolism	3	18	2.47E-04	1.17E-02
Valine, leucine and isoleucine biosynthesis	Fatty acid metabolism	3	6	1.00E-06	1.31E-04
Valine, leucine and isoleucine biosynthesis	Glycine, serine and threonine metabolism	3	9	1.20E-05	8.14E-04
Valine, leucine and isoleucine biosynthesis	Histidine metabolism	6	6	0.00E+00	0.00E+00
Valine, leucine and isoleucine biosynthesis	Lipopolysaccharide biosynthesis	6	12	0.00E+00	0.00E+00
Valine, leucine and isoleucine biosynthesis	Pantothenate and CoA biosynthesis	3	6	1.00E-06	1.31E-04
Valine, leucine and isoleucine biosynthesis	Phenylalanine, tyrosine and tryptophan biosynthesis	3	15	1.15E-04	6.19E-03
Valine, leucine and isoleucine biosynthesis	Porphyrin and chlorophyll metabolism	3	6	1.00E-06	1.31E-04
Valine, leucine and isoleucine biosynthesis	Propanoate metabolism	3	3	0.00E+00	0.00E+00
Valine, leucine and isoleucine biosynthesis	Riboflavin metabolism	3	3	0.00E+00	0.00E+00
Valine, leucine and isoleucine biosynthesis	Thiamine metabolism	3	3	0.00E+00	0.00E+00
Valine, leucine and	Transporters	29	270	0.00E+00	0.00E+00

isoleucine biosynthesis					
Valine, leucine and isoleucine biosynthesis	Vitamin B6 metabolism	3	3	0.00E+00	0.00E+00

Table 7.2: Reduction in number of significantly complemented function pairs in environmental metacommunities is not due to loss of functional richness

	SCFPs	functions	possible	complemented	β-diversity	$p <$
Global	1353	394	45149 (58.3%)	42427 (94.0%)	0.993 (+/- 0.0284)	--
Soil	369	363	28879 (44.0%)	22441 (77.7%)	0.987 (+/- 0.0632)	2.31×10^{-38}
Aquatic	478	375	38278 (54.6%)	31075 (81.2%)	0.988 (+/- 0.0577)	3.39×10^{-65}
Gut	291	374	35520 (50.9%)	9396 (82.8%)	0.976 (+/- 0.116)	1.37×10^{-10}

Table 7.3: Betweenness centrality, distance, convergence time, and chemical similarity of the substrates of SCFPs of three different metacommunities

	Betweenness centrality			Distance		
	$p <$	SCFPs	Other	$p <$	SCFPs	Other
Gut	0.112	0.008	0.009	10^{-300}	7.292	7
Aquatic	10^{-300}	0.006	0.009	1	8.583	7
Soil	0.769	0.009	0.009	10^{-300}	6.5	7

	Convergence time			Chemical similarity		
	$p <$	SCFPs	Other	$p <$	SCFPs	Other
Gut	0.015	4	4.2	10^{-300}	0.435	0.292
Aquatic	0.991	4.69	4.25	10^{-300}	0.336	0.283
Soil	0.228	4	4.241	10^{-300}	0.25	0.287

7.3. Supporting Figures

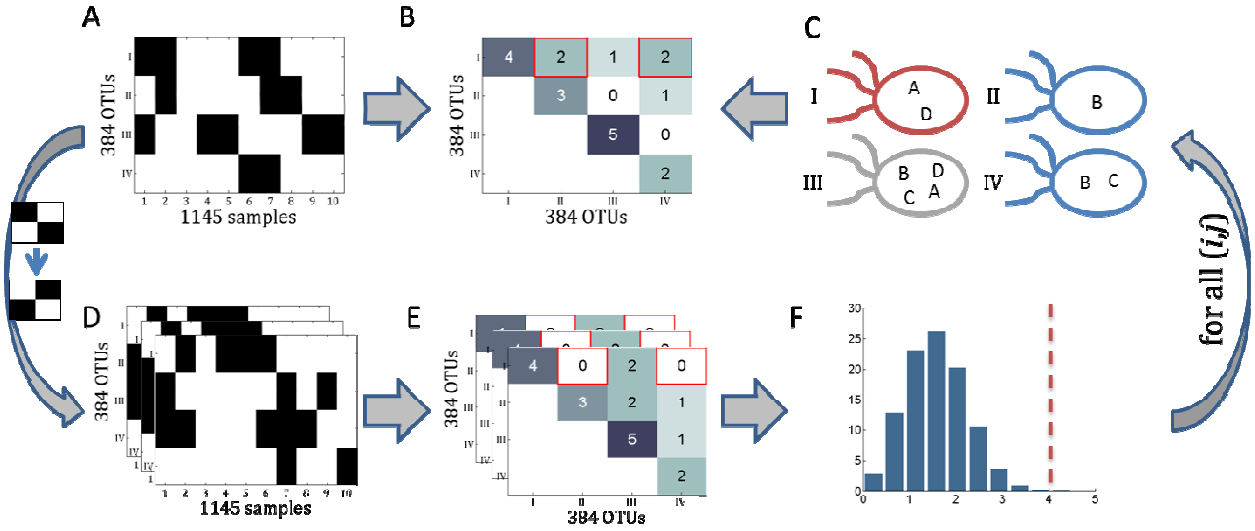


Figure 7.1: An illustration of the method employed to determine significance of functional complements

(A-C) Determining the complementarity score from co-incidence. From the incidence matrix (A), the co-incidence of all OTU pairs is determined (B). From the function annotations (C), a pair (*i*, *j*) is selected: in this example, (A, B) is chosen; OTUs with A but not B are colored red, those with B but not A are colored blue (all others are grey). The co-incidence of red and blue OTUs (highlighted in red boxes) is summed, which in this example yields a complementarity score of 4. (D-F) Null model analysis. The incidence matrix is permuted as described [41,98], and a set of null incidence matrices is generated (D). Co-incidence of target OTUs is summed in each null co-incidence matrix (E). (F) The distribution of null complementarity scores (blue bars) is compared to the experimental score (red dashed line) to assign a *p*-value. Finally, this process is repeated for all function pairs.

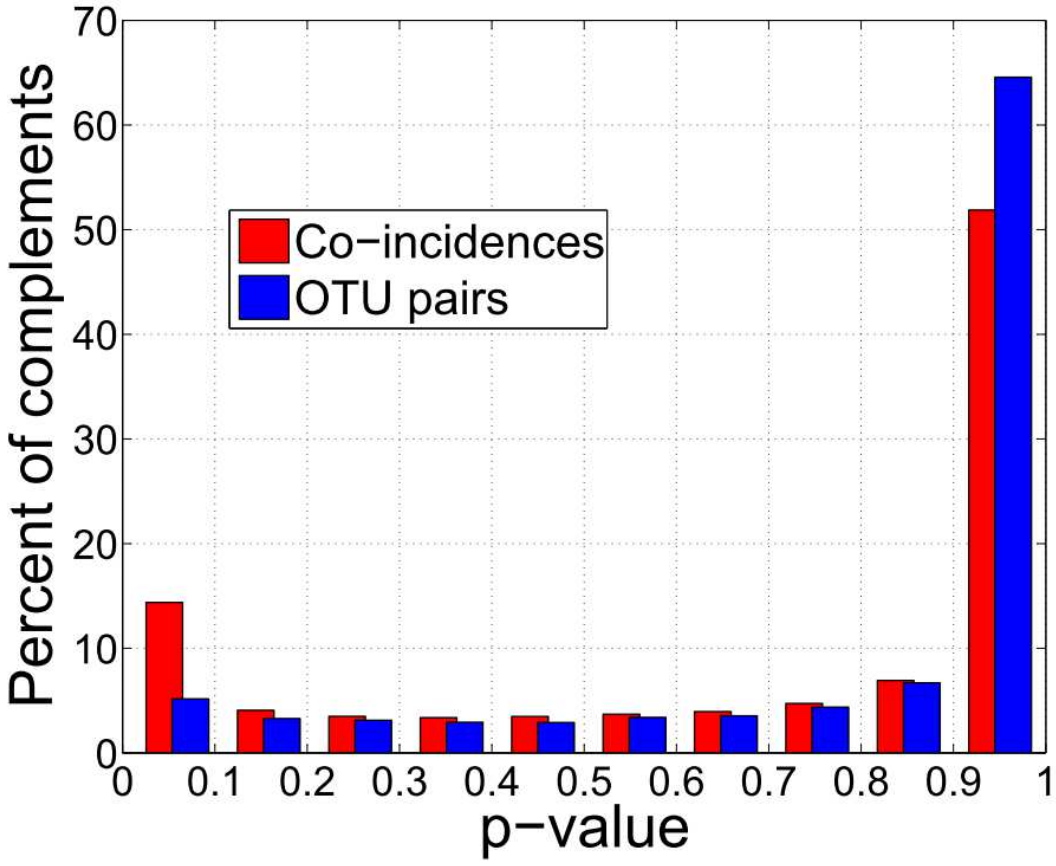


Figure 7.2: Distribution of complementarity p-values

Complementarity scores were calculated in two ways: The number of coincidence events forming a complement (blue) and the number of unique OTU pairs forming a complement (red) (section 7.1.1). The p-value distribution for both scores is heavily skewed such that most complementarities have p-values very close to 1.0, indicating strong tendency away from complementarity in nature. Only pairs of functions which were found complemented at least once are considered here.

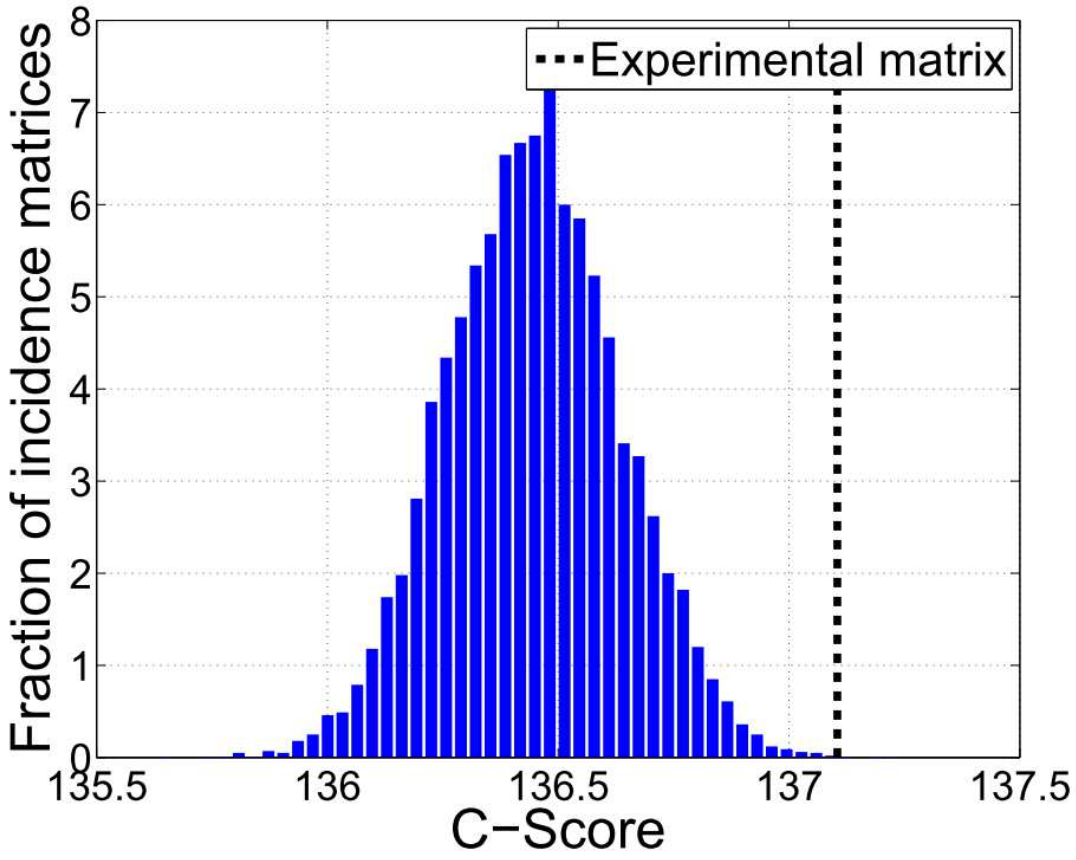


Figure 7.3: Organisms are strongly segregated

The C score of the incidence matrix (dashed line) is shown compared to the distribution of C scores of 10,000 null incidence matrices. Only 4 null matrices approach the degree of checkerboardedness observed in the experimentally determined matrix ($p < 4.9995 \times 10^{-4}$). This is a characteristic feature of populations influenced by species assortment, or more likely in this case, habitat filtering.

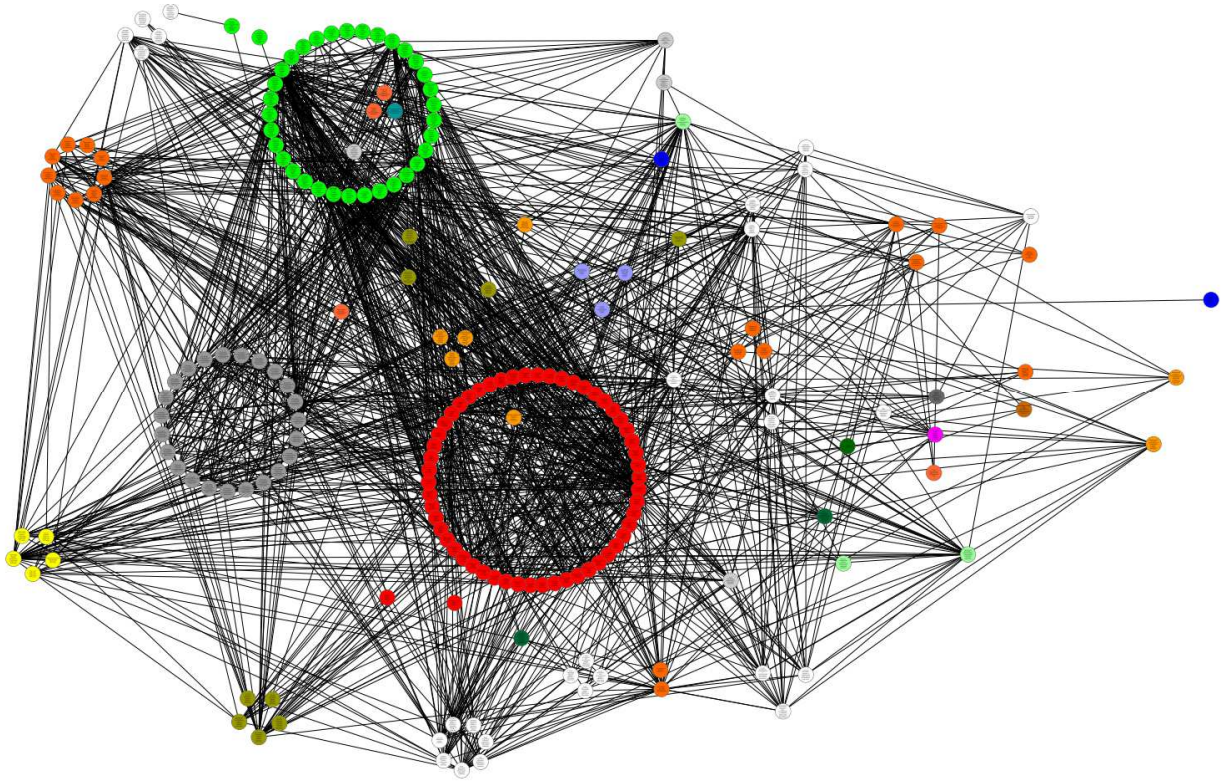


Figure 7.4: A network of significantly complemented function pairs

Each node represents a function and edges are drawn between each pair of complementary functions. Nodes are grouped and colored according to their category (Methods). In total there are 1,325 edges connecting 205 nodes.

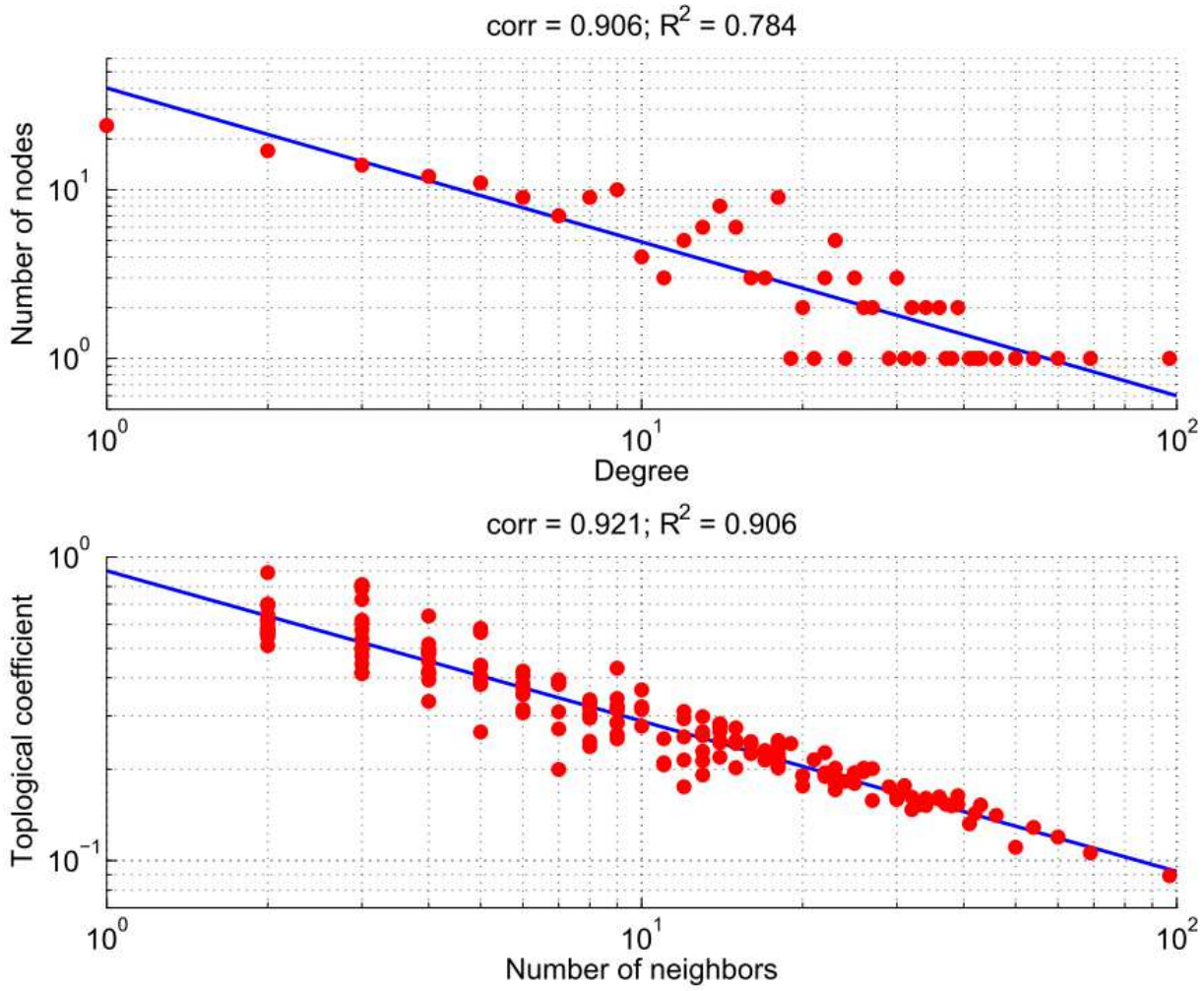


Figure 7.5: Degree and topological coefficient distributions of nodes in the complementarity network

Both approximate a power law, indicating that hub nodes exist but do not cluster together.

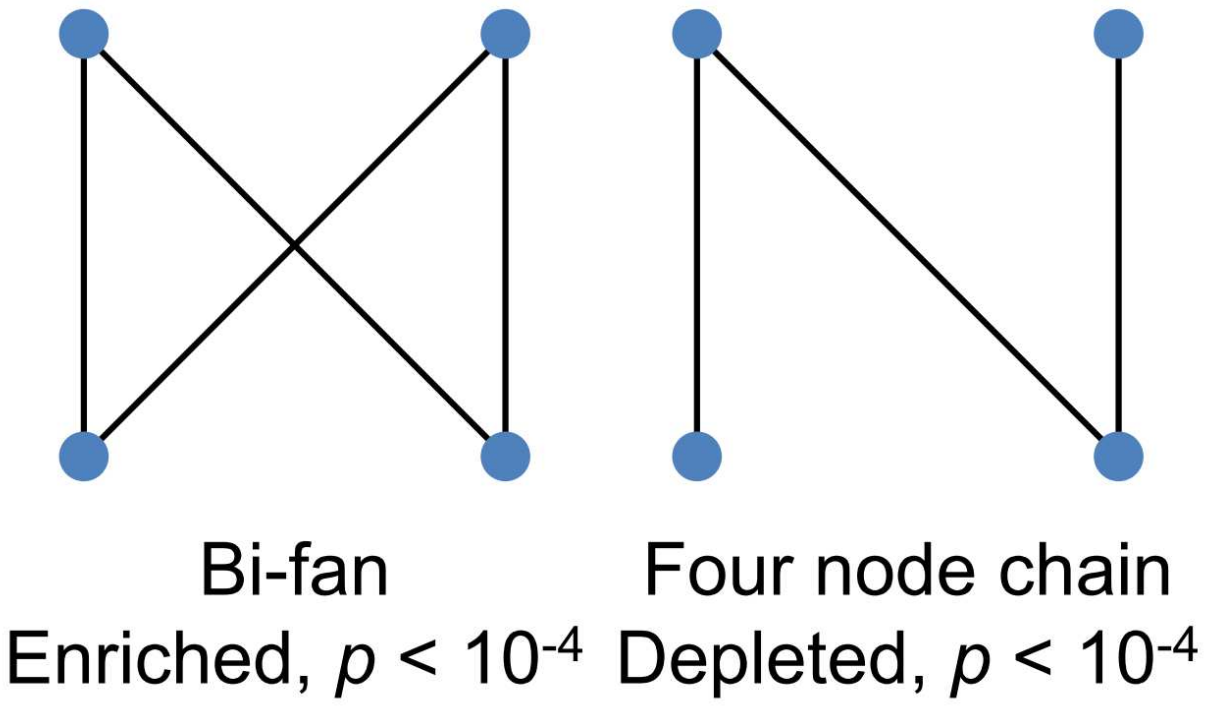
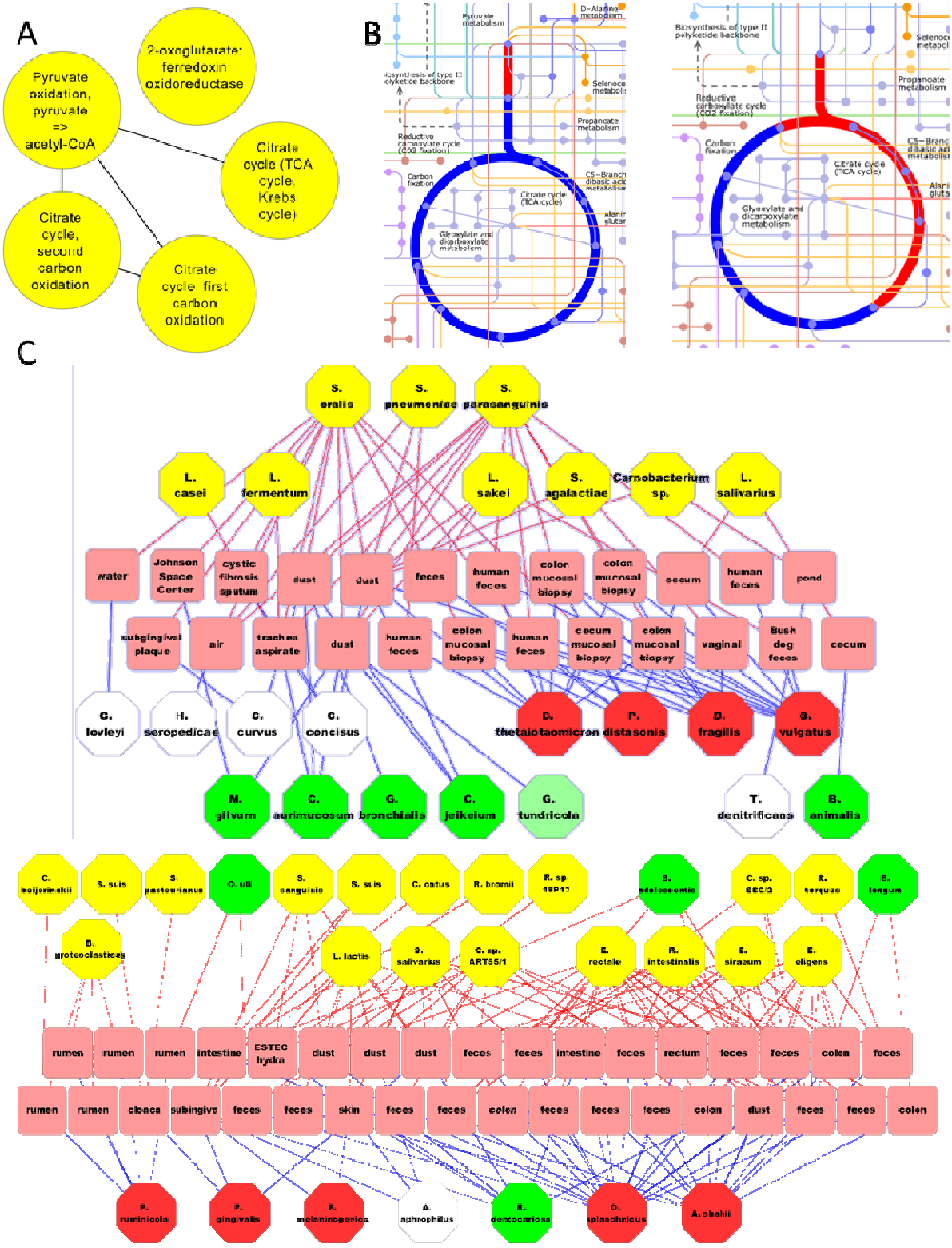


Figure 7.6: A pair of 4-node network motifs enriched and depleted within the SCFP network

The bi-fan motif (left) appears more frequently while the four-node-chain (right) appears less frequently than expected by chance, indicating that functions which share some complements tend to share all.

Figure 7.7 (next page): Complementary functions within the TCA cycle

(A) Two classes of complementarity are found: first, between pyruvate oxidation and the canonical TCA cycle, and between the first and second oxidation step of the TCA cycle. (B) The chemical reactions involved in these complementarities. Left: pyruvate oxidation is shown in red, the TCA cycle in blue. Right: The TCA cycle, with the first carbon oxidation in red and the second in blue. (C) The organisms and samples generating these complements, colored by lineage (yellow: Firmicutes; red: bacteroides; white: proteobacteria; green: actinobacteria). Top: pyruvate oxidation and the TCA cycle; OTUs of the lineage lactobacillales, are known to lack the TCA cycle (except by horizontal transfer). Bottom: first and second carbon oxidation of the TCA cycle. Firmicutes predominantly provide the first oxidation step, Bacteroidetes provide the second.


A
2-oxoglutarate: ferredoxin oxidoreductase
Pyruvate oxidation, pyruvate => acetyl-CoA
Citrate cycle (TCA cycle, Krebs cycle)
Citrate cycle, second carbon oxidation
Citrate cycle, first carbon oxidation
B
C
S. oralis
S. pneumoniae
S. parasanguinis
L. casei
L. fermentum
L. sakei
S. agalactiae
Carnobacterium sp.
L. salivarius
water
Johnson Space Center
cystic fibrosis sputum
dust
dust
feces
human feces
colon mucosal biopsy
colon mucosal biopsy
cecum
human feces
pond
subgingival plaque
air
trachea aspirate
dust
human feces
colon mucosal biopsy
human feces
cecum mucosal biopsy
colon mucosal biopsy
vaginal
Bush dog feces
cecum
G. lovleyi
H. seropedicae
C. curvus
C. concisus
B. thetaiotaomicron
P. distasonis
B. fragilis
B. vulgatus
M. gilvum
C. aurimucosum
G. bronchialis
C. jeikeium
G. fundricola
T. denitrificans
B. animalis
C. beijerinckii
S. suis
S. pastorianus
O. uli
S. sanguinis
S. suis
C. catus
R. bromii
R. sp. 18P13
B. adolescentis
C. sp. SSC/2
R. torques
B. longum
B. proteoclasticus
L. lactis
S. salivarius
C. sp. ART55/1
E. rectale
R. intestinalis
E. siraeum
E. eligens
rumen
rumen
rumen
intestine
ESTEC hydra
dust
dust
dust
feces
feces
intestine
feces
rectum
feces
feces
colon
feces
rumen
rumen
cloaca
subingiva
feces
feces
skin
feces
feces
colon
feces
feces
feces
colon
dust
feces
feces
colon
P. ruminicola
P. gingivalis
F. nucleatum
A. aphrophilus
R. dentocariosa
O. splanchnicus
A. shahii

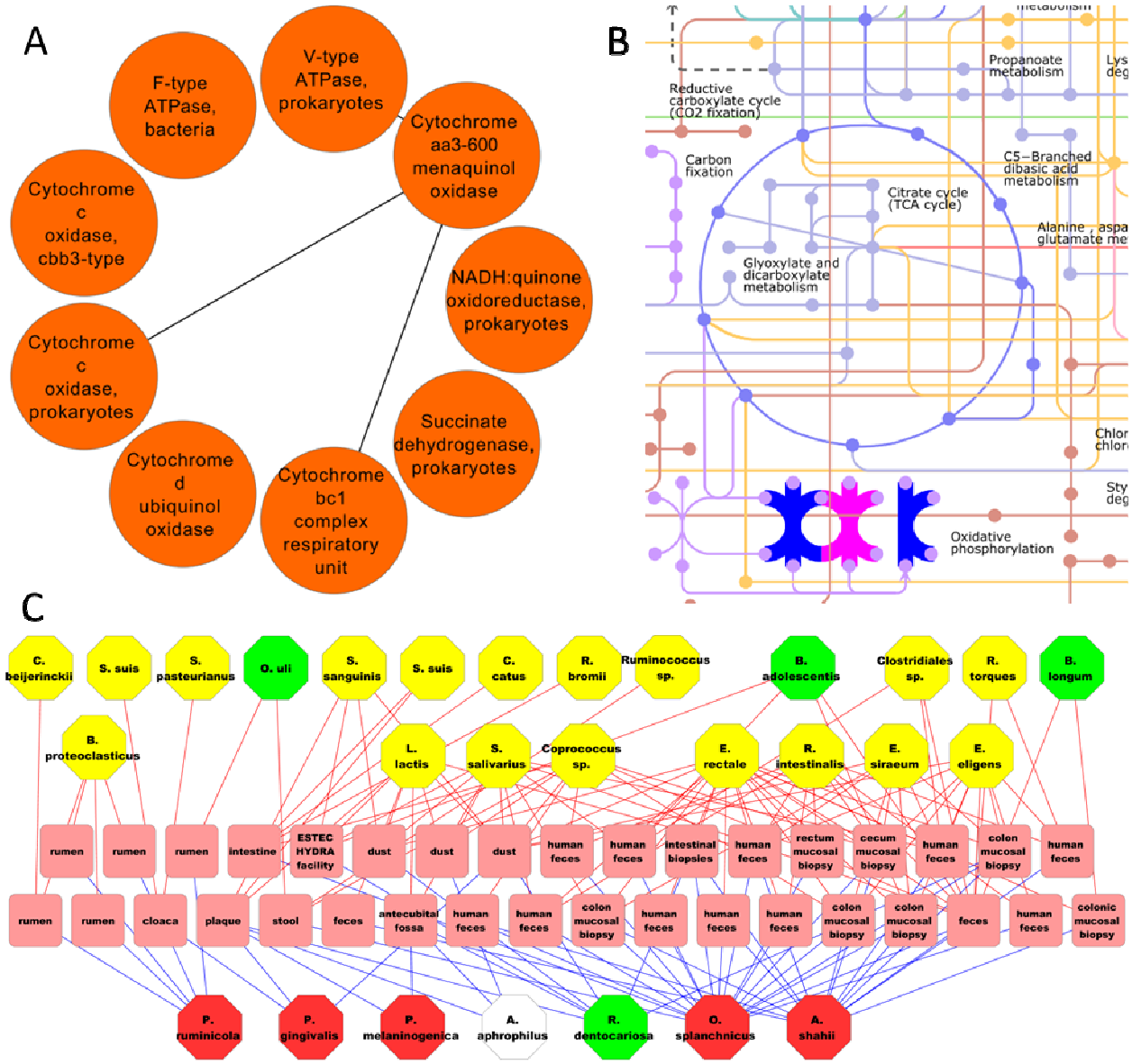


Figure 7.8: Complementary functions within oxidative phosphorylation

(A) All complements are between Cytochrome aa3-600 and one other component of the electron transport chain. (B) The chemical reactions involved in these complementarities. Cytochrome aa3-600 is shown in red, the three other functions in blue (complex IV, represented by Cytochrome aa3-600 and cytochrome c, is shown in purple). (C) The organisms and samples causative of these complements, colored by lineage (yellow: firmicutes; Pink: thermotoga; gray: fusobacteria; purple: deinococci). Bacilli are known to maintain an alternative electron transport chain, shown here to be highly complementary with the canonical form.

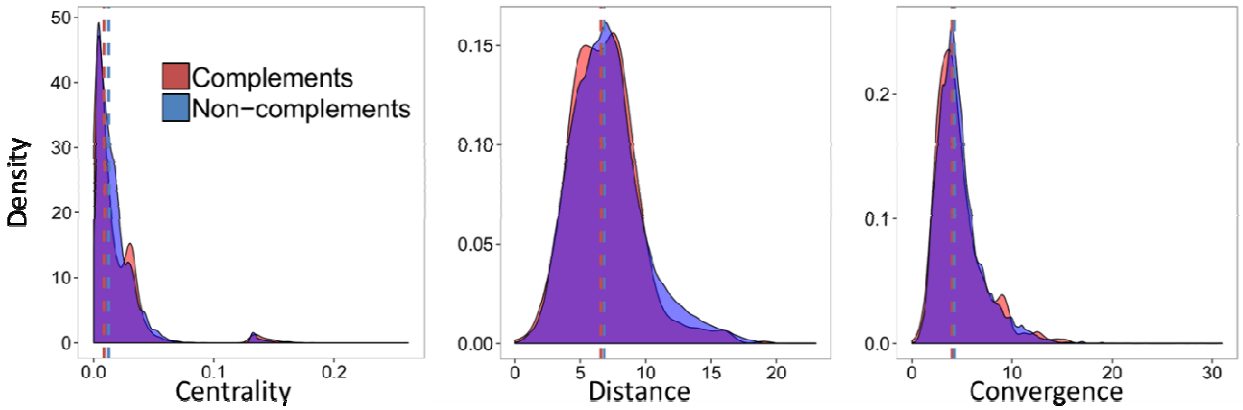


Figure 7.9: Complementarities tend to be found at the periphery of metabolic networks, yet between closely related functions

In all panels, significantly complemented function pairs are shown in red, all others are in blue, and dashed lines correspond to median value: (A) the distributions of mean betweenness centrality, (B) shortest path length, and (C) convergence time of all function pairs. In all cases, the distribution for complementary pairs of functions is significantly lower than for non-complementary functions ($p < 9.01 \times 10^{-9}$, $p < 10^{-300}$, & $p < 0.0395$, respectively).

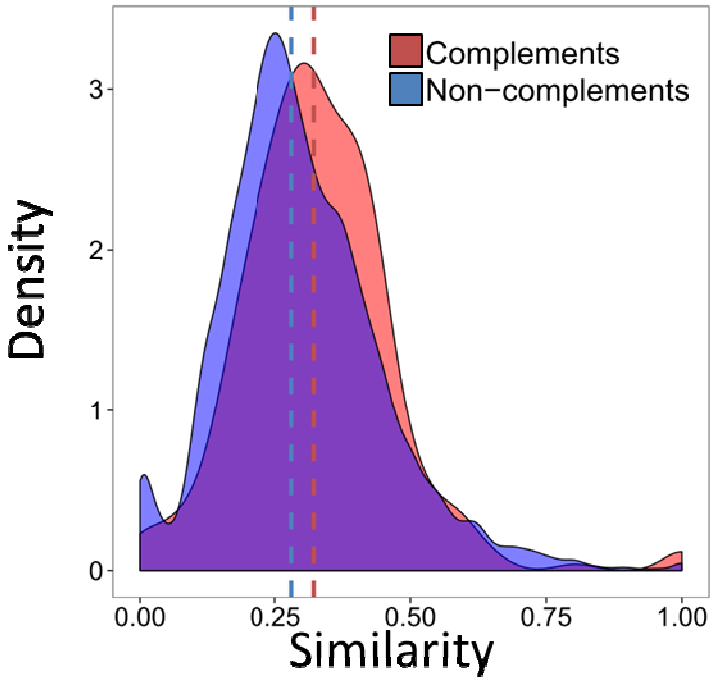


Figure 7.10: Complemented functions act on more chemically similar compounds
Chemical similarity is calculated as the Tanimoto coefficient from common atomic substructure. Complementary functions are shown in red, non-complementary functions in blue. The distribution for complementary pairs of functions is significantly greater than for non-complementary functions.

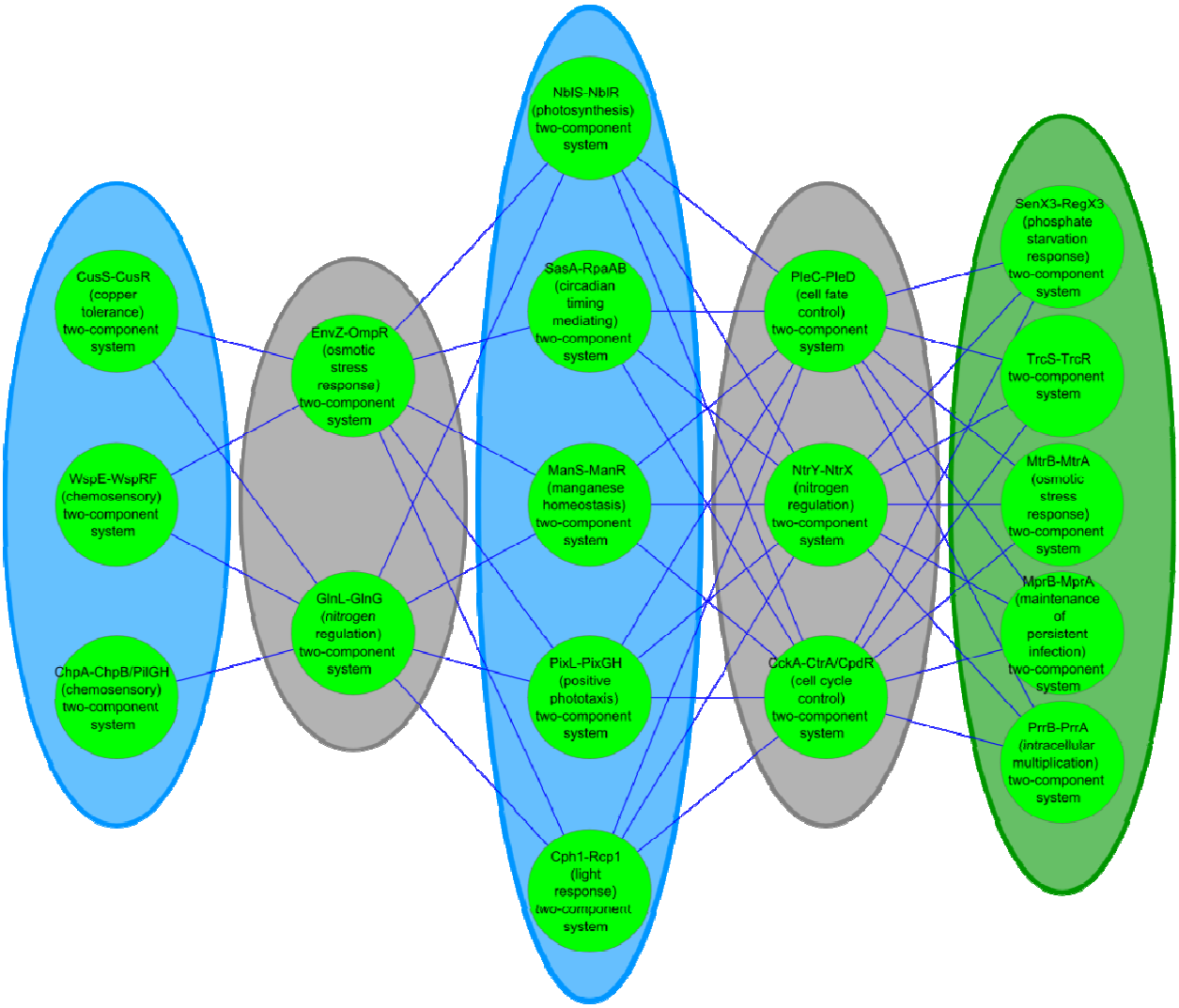


Figure 7.11: Significantly complemented two-component systems in the soil metacommunity

Only the connected component including photosynthesis and nitrogen fixation systems are shown. Bands indicate sets of functions contributed predominantly by taxa of a given clade (green: actinobacteria; grey: proteobacteria, esp. rhizobiales; blue: cyanobacteria, esp. *Nostoc punctiforme*).

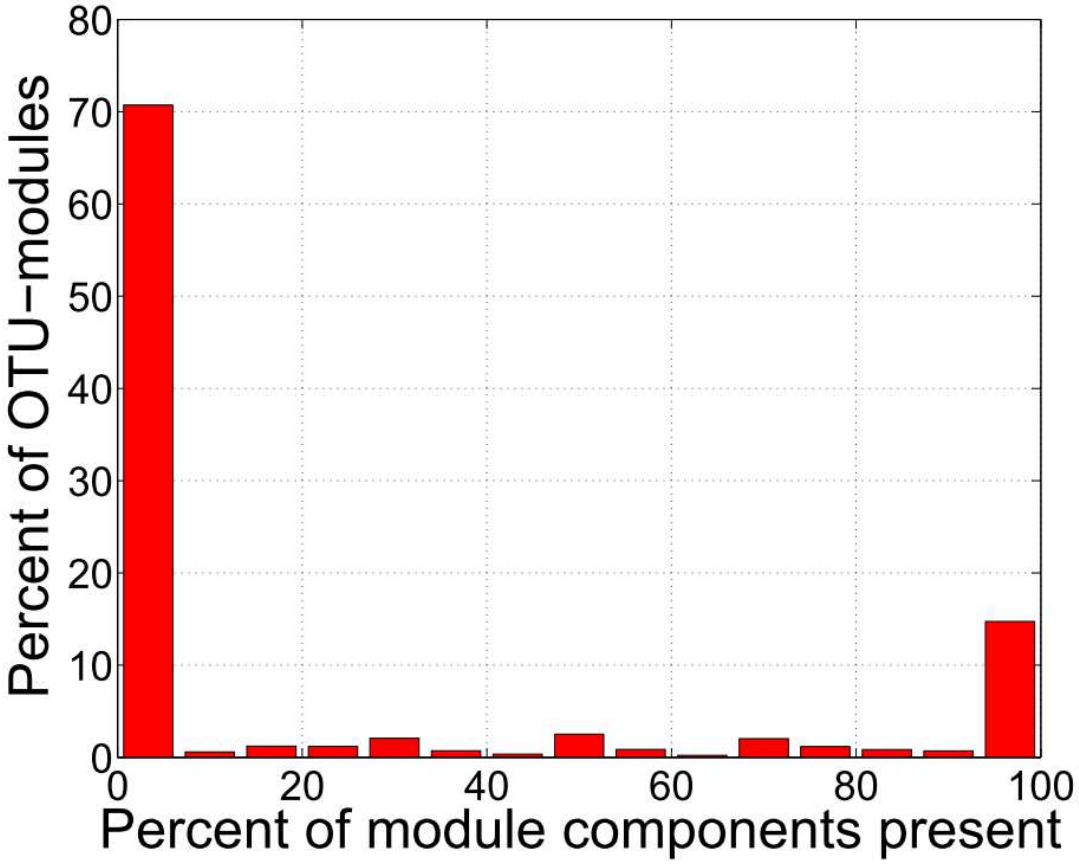


Figure 7.12: The distribution of module components across genomes

The vast majority of modules are either completely present or completely absent from genomes; altering the upper and lower threshold for module presence across genomes does not significantly alter genome annotation.

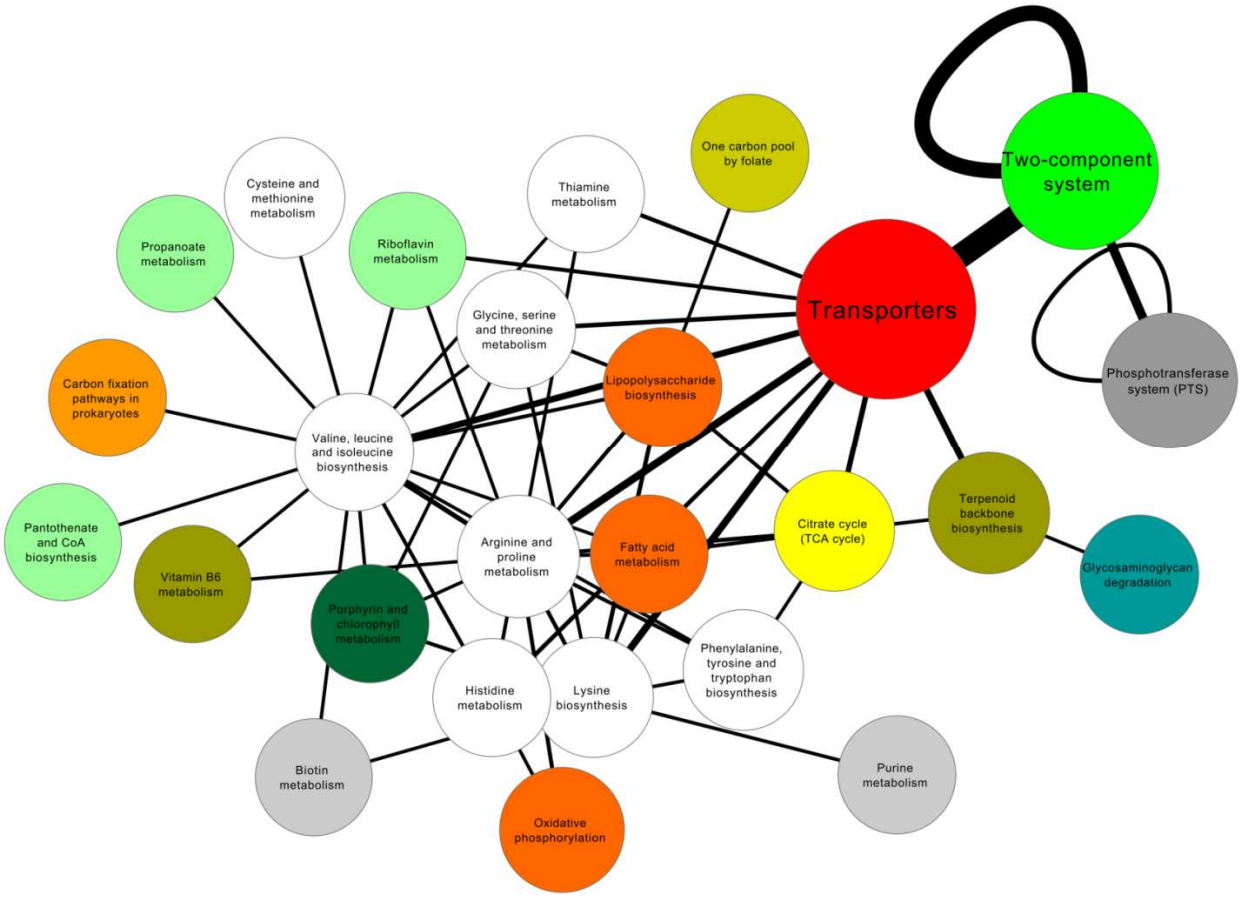


Figure 7.13: Summary network of functional complementary describing the alternative method of scoring complements

Functions are aggregated according to their category, and node size is scaled according to the number of functions. The largest 3 nodes represent transporters, two-component systems, and phosphotransferase systems (with 73, 47, and 22 functions each, respectively). Edges connect categories with a significant number of complementarities after FDR correction (FDR < 5%). Edge sizes are scaled by the number of complementarities between (or within) a given category (with the greatest number of edges between transporters and two-component systems, 183). Only interactions supported by at least 3 complementary function pairs are shown. See also Table 7.1.

8. Appendix C: Supporting Information for Chapter 4

8.1. Supporting Figures

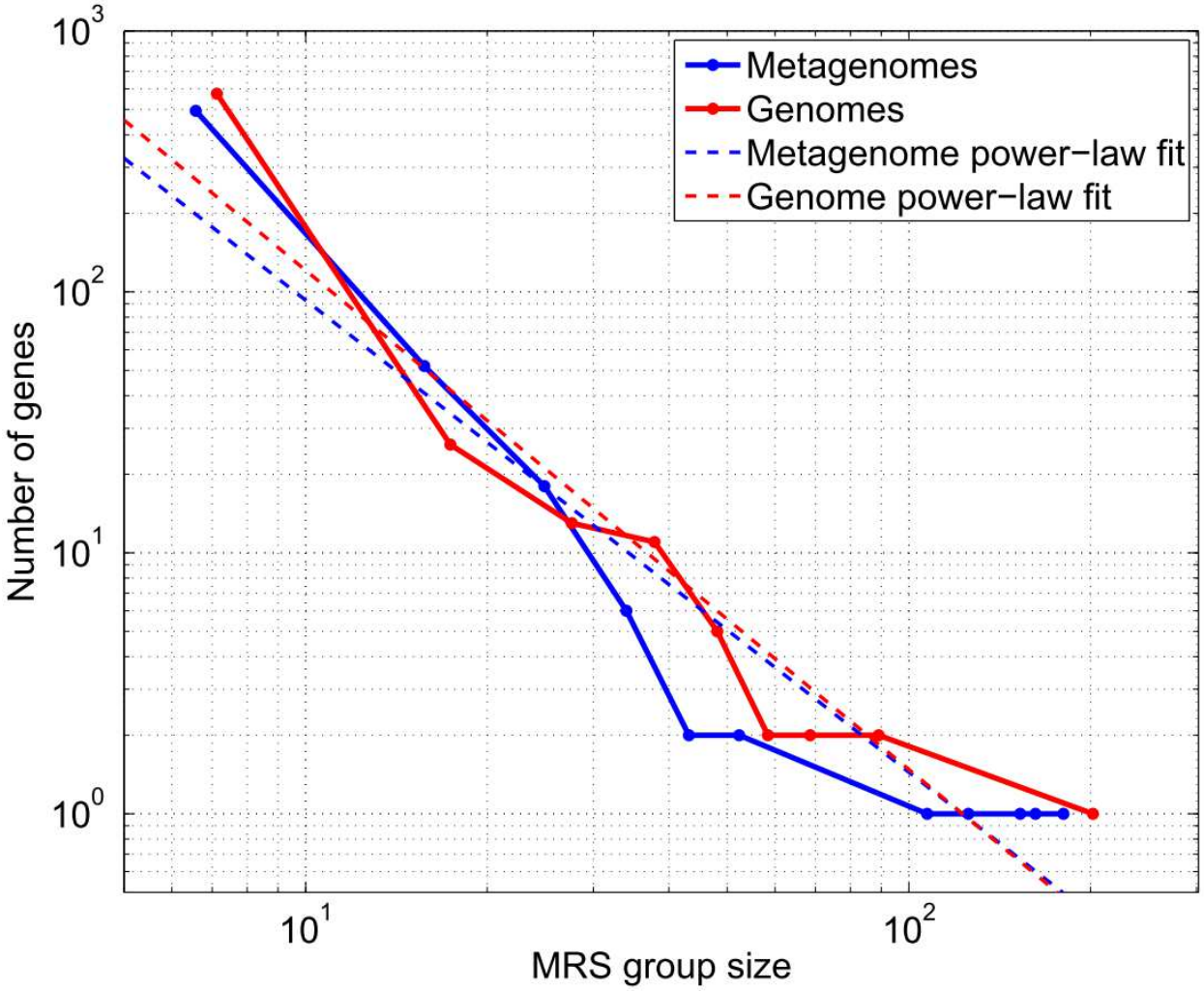


Figure 8.1: MRS component size distributions follow a power law

Solid lines represent the component size distributions of the gene co-occurrence MRSs formed from metagenomes (blue) and genomes (red). Dashed lines represent the fit of each to a power-law.

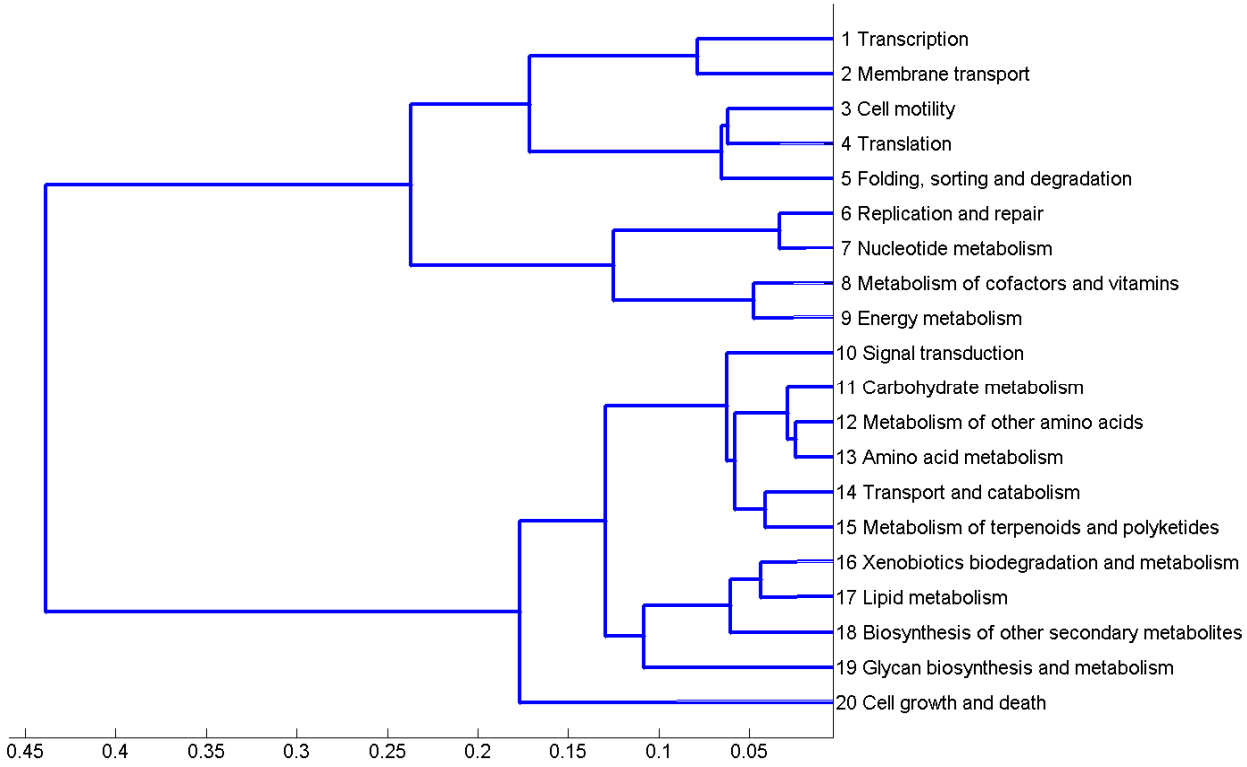


Figure 8.2: Functions cluster based on differential coordination in metagenomes and genomes

Functions in the upper cluster (categories 1 through 9) tend to have greater inter-pathway co-occurrence in genomes, and concern core cellular processes such as genetic information processing. Conversely, the lower cluster (categories 10 through 20) tends to have greater inter-pathway co-occurrence in metagenomes, and includes functions involved in niche selection, such as xenobiotics degradation and secondary metabolism.

9. References

1. Luckey TD (1972) Introduction to intestinal microecology. Am J Clin Nutr 25: 1292–1294.

2. Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, et al. (2012) Structure, function and diversity of the healthy human microbiome. Nature 486: 207–214. doi:10.1038/nature11234.

3. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, et al. (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464: 59–65. doi:10.1038/nature08821.

4. Dehority BA (1991) Effects of microbial synergism on fibre digestion in the rumen. Proc Nutr Soc 50: 149–159.

5. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, et al. (2006) Metagenomic analysis of the human distal gut microbiome. Science 312: 1355–1359. doi:10.1126/science.1124234.

6. Robert C, Bernalier-Donadille A (2003) The cellulolytic microflora of the human colon: evidence of microcrystalline cellulose-degrading bacteria in methane-excreting subjects. FEMS Microbiol Ecol 46: 81–89. doi:10.1016/S0168-6496(03)00207-1.

7. Wilson ID, Nicholson JK (2009) The role of gut microbiota in drug response. Curr Pharm Des 15: 1519–1523.

8. Upreti RK, Shrivastava R, Chaturvedi UC (2004) Gut microflora & toxic metals: chromium as a model. Indian J Med Res 119: 49–59.

9. Boxenbaum HG, Bekersky I, Jack ML, Kaplan SA (1979) Influence of gut microflora on bioavailability. Drug Metab Rev 9: 259–279. doi:10.3109/03602537908993894.

10. Illing HP (1981) Techniques for microfloral and associated metabolic studies in relation to the absorption and enterohepatic circulation of drugs. Xenobiotica 11: 815–830. doi:10.3109/00498258109045319.

11. Lee SM, Donaldson GP, Mikulski Z, Boyajian S, Ley K, et al. (2013) Bacterial colonization factors control specificity and stability of the gut microbiota. Nature 501: 426–429. doi:10.1038/nature12447.

12. Round JL, Mazmanian SK (2009) The gut microbiota shapes intestinal immune responses during health and disease. Nat Rev Immunol 9: 313–323.

13. Kelly D, Campbell JI, King TP, Grant G, Jansson E a, et al. (2004) Commensal anaerobic gut bacteria attenuate inflammation by regulating nuclear-cytoplasmic shuttling of PPAR-gamma and RelA. Nat Immunol 5: 104–112. doi:10.1038/ni1018.

14. Stolyar S, Van Dien S, Hillesland KL, Pinel N, Lie TJ, et al. (2007) Metabolic modeling of a mutualistic microbial community. Mol Syst Biol 3: 92. doi:10.1038/msb4100131.

15. Miller TL, Wolin MJ, de Macario EC, Macario AJ (1982) Isolation of Methanobrevibacter smithii from human feces. Appl Envir Microbiol 43: 227–232.

16. Rakoff-Nahoum S, Coyne MJ, Comstock LE (2014) An ecological network of polysaccharide utilization among human intestinal symbionts. Curr Biol 24: 40–49. doi:10.1016/j.cub.2013.10.077.

17. Mahowald MA, Rey FE, Seedorf H, Turnbaugh PJ, Fulton RS, et al. (2009) Characterizing a model human gut microbiota composed of members of its two dominant bacterial phyla. Proc Natl Acad Sci U S A 106: 5859–5864. doi:10.1073/pnas.0901529106.

18. Ley RE, Bäckhed F, Turnbaugh P, Lozupone C a, Knight RD, et al. (2005) Obesity alters gut microbial ecology. Proc Natl Acad Sci U S A 102: 11070–11075. doi:10.1073/pnas.0504978102.

19. Schwiertz A, Taras D, Schäfer K, Beijer S, Bos NA, et al. (2010) Microbiota and SCFA in lean and overweight healthy subjects. Obesity (Silver Spring) 18: 190–195. doi:10.1038/oby.2009.167.

20. Ridaura VK, Faith JJ, Rey FE, Cheng J, Duncan AE, et al. (2013) Gut microbiota from twins discordant for obesity modulate metabolism in mice. Science 341: 1241214. doi:10.1126/science.1241214.

21. Øvreås L, Daae FL, Torsvik V, Rodríguez-Valera F (2003) Characterization of microbial diversity in hypersaline environments by melting profiles and reassociation kinetics in combination with terminal restriction fragment length polymorphism (T-RFLP). Microb Ecol 46: 291–301. doi:10.1007/s00248-003-3006-3.

22. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, et al. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428: 37–43.

23. Curtis TP, Sloan WT, Scannell JW (2002) Estimating prokaryotic diversity and its limits. Proc Natl Acad Sci U S A 99: 10494–10499. doi:10.1073/pnas.142680199.

24. Torsvik V, Salte K, Sørheim R, Goksøyr J (1990) Comparison of phenotypic diversity and DNA heterogeneity in a population of soil bacteria. Appl Environ Microbiol 56: 776–781.

25. Torsvik V, Øvreås L, Thingstad TF (2002) Prokaryotic diversity--magnitude, dynamics, and controlling factors. Science 296: 1064–1066. doi:10.1126/science.1071698.

26. Chaffron S, Rehrauer H, Pernthaler J, von Mering C (2010) A global network of coexisting microbes from environmental and whole-genome sequence data. Genome Res 20: 947–959. doi:10.1101/gr.104521.109.

27. Boetius A, Ravenschlag K, Schubert CJ, Rickert D, Widdel F, et al. (2000) A marine microbial consortium apparently mediating anaerobic oxidation of methane. Nature 407: 623–626.

28. Knoll AH (2003) The geological consequences of evolution. Geobiology 1: 3–14. doi:10.1046/j.1472-4669.2003.00002.x.

29. Parveen B, Ravet V, Djediat C, Mary I, Quiblier C, et al. (2013) Bacterial communities associated with Microcystis colonies differ from free-living communities living in the same ecosystem. Environ Microbiol Rep 5: 716–724. doi:10.1111/1758-2229.12071.

30. Philippot L, Spor A, Hénault C, Bru D, Bizouard F, et al. (2013) Loss in microbial diversity affects nitrogen cycling in soil. ISME J 7: 1609–1619. doi:10.1038/ismej.2013.34.

31. Staley JT, Konopka A (1985) Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annu Rev Microbiol 39: 321–346. doi:10.1146/annurev.mi.39.100185.001541.

32. Winker S, Woese CR (1991) A definition of the domains Archaea, Bacteria and Eucarya in terms of small subunit ribosomal RNA characteristics. Syst Appl Microbiol 14: 305–310. doi:10.1016/S0723-2020(11)80303-6.

33. Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, et al. (1985) Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proc Natl Acad Sci U S A 82: 6955–6959.

34. Schmidt TM, DeLong EF, Pace NR (1991) Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. J Bacteriol 173: 4371–4378.

35. Bik EM, Long CD, Armitage GC, Loomer P, Emerson J, et al. (2010) Bacterial diversity in the oral cavity of 10 healthy individuals. ISME J 4: 962–974. doi:10.1038/ismej.2010.30.

36. Horner-Devine MC, Silver JM, Leibold M a, Bohannan BJM, Colwell RK, et al. (2007) A comparison of taxon co-occurrence patterns for macro- and microorganisms. Ecology 88: 1345–1353.

37. Jeraldo P, Sipos M, Chia N, Brulca JM, Dhillond AS, et al. (2012) Quantification of the relative roles of niche and neutral processes in structuring gastrointestinal microbiomes. Proc Natl Acad Sci U S A 109: 9692–9698. doi:10.1073/pnas.1206721109/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1206721109.

38. Zhou J, Xia B, Treves DS, Wu L-Y, Marsh TL, et al. (2002) Spatial and Resource Factors Influencing High Microbial Diversity in Soil. Appl Environ Microbiol 68: 326–334. doi:10.1128/AEM.68.1.326-334.2002.

39. Horner-Devine MC, Bohannan BJM (2006) Phylogenetic clustering and overdispersion in bacterial communities. Ecology 87: S100–S108.

40. Cody M, Diamond J (1975) Ecology and Evolution of Communities (Belknap Press). Belknap Press of Harvard University Press.

41. Connor EF, Simberloff D (1979) The Assembly of Species Communities: Chance or Competition? Ecology 60: 1132. doi:10.2307/1936961.

42. Mee MT, Collins JJ, Church GM, Wang HH (2014) Syntrophic exchange in synthetic microbial communities. Proc Natl Acad Sci 111: E2149–E2156. doi:10.1073/pnas.1405641111.

43. Kim P-J, Price ND (2011) Genetic co-occurrence network across sequenced microbes. PLoS Comput Biol 7: e1002340. doi:10.1371/journal.pcbi.1002340.

44. Joseph Felsenstein (1985) Phylogenies and the comparative method. Am Nat 125: 1–15.

45. Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, et al. (2009) Bacterial community variation in human body habitats across space and time. Science 326: 1694–1697. doi:10.1126/science.1177486.

46. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, et al. (2009) A core gut microbiome in obese and lean twins. Nature 457: 480–484. doi:10.1038/nature07540.

47. Ley RE (2010) Obesity and the human microbiome. Curr Opin Gastroenterol 26: 5–11.

48. Qin J, Li Y, Cai Z, Li S, Zhu J, et al. (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 490: 55–60. doi:10.1038/nature11450.

49. Faith JJ, McNulty NP, Rey FE, Gordon JI (2011) Predicting a Human Gut Microbiota's Response to Diet in Gnotobiotic Mice. Science (80-) 333: 101–104. doi:10.1126/science.1206025.

50. Koenig J, Spor A (2011) Succession of microbial consortia in the developing infant gut microbiome. Proc Natl Acad Sci U S A 108: 4578–4585. doi:10.1073/pnas.1000081107.

51. Emerson BC, Gillespie RG (2008) Phylogenetic analysis of community assembly and structure over space and time. Trends Ecol Evol 23: 619–630. doi:10.1016/j.tree.2008.07.005.

52. Weiher E, Clarke PGD, Keddy PA (1998) Community assembly rules, morphological dispersion, and the coexistence of plant species. Oikos 81: 309–322.

53. Cornwell WK, Schwilk LDW, Ackerly DD (2006) A trait-based test for habitat filtering: convex hull volume. Ecology 87: 1465–1471.

54. Borenstein E, Kupiec M, Feldman MW, Ruppin E (2008) Large-scale reconstruction and phylogenetic analysis of metabolic environments. Proc Natl Acad Sci U S A 105: 14482–14487. doi:10.1073/pnas.0806162105.

55. Janga SC, Babu MM (2008) Network-based approaches for linking metabolism with environment. Genome Biol 9: 239. doi:10.1186/gb-2008-9-11-239.

56. Levy R, Borenstein E (2012) Reverse ecology: from systems to environments and back. Adv Exp Med Biol 751: 329–345.

57. Freilich S, Kreimer A, Borenstein E, Yosef N, Sharan R, et al. (2009) Metabolic-network-driven analysis of bacterial ecological strategies. Genome Biol 10: R61. doi:10.1186/gb-2009-10-6-r61.

58. Borenstein E (2012) Computational systems biology and in silico modeling of the human microbiome. Brief Bioinform 13: 769–780.

59. Borenstein E, Feldman MW (2009) Topological signatures of species interactions in metabolic networks. J Comput Biol 16: 191–200. doi:10.1089/cmb.2008.06TT.

60. Freilich S, Kreimer A, Meilijson I, Gophna U, Sharan R, et al. (2010) The large-scale organization of the bacterial network of ecological co-occurrence interactions. Nucleic Acids Res 38: 3857–3868. doi:10.1093/nar/gkq118.

61. Cottret L, Milreu PV, Acuña V, Marchetti-Spaccamela A, Stougie L, et al. (2010) Graph-based analysis of the metabolic exchanges between two co-resident intracellular symbionts, Baumannia cicadellinicola and Sulcia muelleri, with their insect host,

Homalodisca coagulata. PLoS Comput Biol 6: e1000904. doi:10.1371/journal.pcbi.1000904.

62. Greenblum S, Turnbaugh PJ, Borenstein E (2012) Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. Proc Natl Acad Sci U S A. doi:10.1073/pnas.1116053109/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1116053109.

63. Kolenbrander PE (2011) Multispecies communities: interspecies interactions influence growth on saliva as sole nutritional source. Int J Oral Sci 3: 49–54. doi:10.4248/IJOS11025.

64. Kolenbrander PE, Palmer RJ, Periasamy S, Jakubovics NS (2010) Oral multispecies biofilm development and the key role of cell-cell distance. Nat Rev Microbiol 8: 471–480. doi:10.1038/nrmicro2381.

65. Palmer RJ, Kazmerzak K, Hansen MC, Kolenbrander PE (2001) Mutualism versus independence: strategies of mixed-species oral biofilms in vitro using saliva as the sole nutrient source. Infect Immun 69: 5794–5804. doi:10.1128/IAI.69.9.5794.

66. Periasamy S, Kolenbrander PE (2009) Mutualistic biofilm communities develop with Porphyromonas gingivalis and initial, early, and late colonizers of enamel. J Bacteriol 191: 6804–6811.

67. Markowitz VM, Chen I-M a, Palaniappan K, Chu K, Szeto E, et al. (2012) IMG: the Integrated Microbial Genomes database and comparative analysis system. Nucleic Acids Res 40: D115–D122. doi:10.1093/nar/gkr1044.

68. Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, et al. (2012) Microbial co-occurrence relationships in the human microbiome. PLoS Comput Biol 8: e1002606. doi:10.1371/journal.pcbi.1002606.

69. Zaneveld JR, Lozupone C, Gordon JI, Knight R (2010) Ribosomal RNA diversity predicts genome diversity in gut bacteria and their relatives. Nucleic Acids Res 38: 3869.

70. Louis P, Scott KP, Duncan SH, Flint HJ (2007) Understanding the effects of diet on bacterial metabolism in the large intestine. J Appl Microbiol 102: 1197–1208. doi:10.1111/j.1365-2672.2007.03322.x.

71. Foster KR, Bell T (2012) Competition, Not Cooperation, Dominates Interactions among Culturable Microbial Species. Curr Biol 22: 1845–1850. doi:10.1016/j.cub.2012.08.005.

72. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen Y-Y, et al. (2011) Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes. Science (80-) 105: 105–108. doi:10.1126/science.1208344.

73. Kolida S, Meyer D, Gibson GR (2007) A double-blind placebo-controlled study to establish the bifidogenic dose of inulin in healthy humans. Eur J Clin Nutr 61: 1189–1195. doi:10.1038/sj.ejcn.1602636.

74. Klitgord N, Segrè D (2010) Environments that induce synthetic microbial ecosystems. PLoS Comput Biol 6: e1001002. doi:10.1371/Citation.

75. Cordero OX, Wildschutte H, Kirkup B, Proehl S, Ngo L, et al. (2012) Ecological populations of bacteria act as socially cohesive units of antibiotic production and resistance. Science 337: 1228–1231. doi:10.1126/science.1219385.

76. Turnbaugh PJ, Gordon JI (2008) An invitation to the marriage of metagenomics and metabolomics. Cell 134: 708–713. doi:10.1016/j.cell.2008.08.025.

77. Greenblum S, Chiu H-C, Levy R, Carr R, Borenstein E (2013) Towards a predictive systems-level model of the human microbiome: progress, challenges, and opportunities. Curr Opin Biotechnol 24: 1–11. doi:10.1016/j.copbio.2013.04.001.

78. Taormina MJ, Jemielita M, Stephens WZ, Burns AR, Troll J V, et al. (2012) Investigating Bacterial-Animal Symbioses with Light Sheet Microscopy. Biol Bull 223: 7–20.

79. Lemon KP, Armitage GC, Relman D a., Fischbach M a. (2012) Microbiota-targeted therapies: an ecological perspective. Sci Transl Med 4: 137rv5. doi:10.1126/scitranslmed.3004183.

80. Röling WFM, Ferrer M, Golyshin PN (2010) Systems approaches to microbial communities and their functioning. Curr Opin Biotechnol 21: 532–538. doi:10.1016/j.copbio.2010.06.007.

81. Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 28: 27–30.

82. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, et al. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72: 5069–5072. doi:10.1128/AEM.03006-05.

83. Segata N, Waldron L, Ballarini A (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods 9: 811–814. doi:10.1038/Nmeth.2066.

84. Fierer N, Jackson RB (2006) The diversity and biogeography of soil bacterial communities. Proc Natl Acad Sci U S A 103: 626–631. doi:10.1073/pnas.0507535103.

85. Barberán A, Bates ST, Casamayor EO, Fierer N (2011) Using network analysis to explore co-occurrence patterns in soil microbial communities. ISME J: 1–9. doi:10.1038/ismej.2011.119.

86. Woodcock S, van der Gast CJ, Bell T, Lunn M, Curtis TP, et al. (2007) Neutral assembly of bacterial communities. FEMS Microbiol Ecol 62: 171–180. doi:10.1111/j.1574-6941.2007.00379.x.

87. Shou W, Ram S, Vilar JMG (2007) Synthetic cooperation in engineered yeast populations. Proc Natl Acad Sci U S A 104: 1877–1882. doi:10.1073/pnas.0610575104.

88. Momeni B, Chen C-C, Hillesland KL, Waite A, Shou W (2011) Using artificial systems to explore the ecology and evolution of symbioses. Cell Mol Life Sci 68: 1353–1368. doi:10.1007/s00018-011-0649-y.

89. Wintermute EH, Silver P a (2010) Emergent cooperation in microbial metabolism. Mol Syst Biol 6: 407. doi:10.1038/msb.2010.66.

90. Shoaie S, Karlsson F, Mardinoglu A, Nookaew I, Bordel S, et al. (2013) Understanding the interactions between bacteria in the human gut through metabolic modeling. Sci Rep 3: 1–10. doi:10.1038/srep02532.

91. Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, et al. (2011) Ecology drives a global network of gene exchange connecting the human microbiome. Nature 480: 241–244. doi:10.1038/nature10571.

92. Karr JRR, Sanghvi JCC, Macklin DNN, Gutschow MV V, Jacobs JMM, et al. (2012) A whole-cell computational model predicts phenotype from genotype. Cell 150: 389–401. doi:10.1016/j.cell.2012.05.044.

93. Edwards JS, Palsson BO (2000) Metabolic flux balance analysis and the in silico analysis of Escherichia coli K-12 gene deletions. BMC Bioinformatics 1: 1.

94. Chiu HH-C, Levy R, Borenstein E (2014) Emergent biosynthetic capacity in simple microbial communities. PLoS Comput Biol 10: e1003695. doi:10.1371/journal.pcbi.1003695.

95. Harcombe W, Riehl W, Dukovski I (2014) Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics. Cell Rep: 1–12. doi:10.1016/j.celrep.2014.03.070.

96. Levy R, Borenstein E (2013) Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules. Proc Natl Acad Sci U S A 110: 12804–12809. doi:10.1073/pnas.1300926110/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1300926110.

97. Levy R, Borenstein E (2014) Metagenomic systems biology and metabolic modeling of the human microbiome. Gut Microbes: 1–6.

98. Gotelli NJ (2000) Null Model Analysis of Species Co-Occurrence Patterns. Ecology 81: 2606–2621. doi:10.1890/0012-9658(2000)081[2606:NMAOSC]2.0.CO;2.

99. Stone L, Roberts A (1990) The checkerboard score and species distributions. Oecologia 85: 74–79. doi:10.1007/BF00317345.

100. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, et al. (2005) A human protein-protein interaction network: a resource for annotating the proteome. Cell 122: 957–968. doi:10.1016/j.cell.2005.08.029.

101. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, et al. (2002) Network motifs: simple building blocks of complex networks. Science 298: 824–827. doi:10.1126/science.298.5594.824.

102. Wu D, Daugherty SC, Van Aken SE, Pai GH, Watkins KL, et al. (2006) Metabolic complementarity and genomics of the dual bacterial symbiosis of sharpshooters. PLoS Biol 4: e188. doi:10.1371/journal.pbio.0040188.

103. Morishita T, Yajima M (1995) Incomplete operation of biosynthetic and bioenergetic functions of the citric acid cycle in multiple auxotrophic lactobacilli. Biosci Biotechnol Biochem 59: 251–255.

104. Lauraeus M, Wikström M (1993) The terminal quinol oxidases of Bacillus subtilis have different energy conservation properties. J Biol Chem 268: 11470–11473.

105. Lemma E, Simon J, Schägger H, Kröger A (1995) Properties of the menaquinol oxidase (Qox) and of qox deletion mutants of Bacillus subtilis. Arch Microbiol 163: 432–438.

106. Yi SM, Narasimhulu K V, Samoilova RI, Gennis RB, Dikanov S a (2010) Characterization of the semiquinone radical stabilized by the cytochrome aa3-600 menaquinol oxidase of Bacillus subtilis. J Biol Chem 285: 18241–18251. doi:10.1074/jbc.M110.116186.

107. Rahman SA, Bashton M, Holliday GL, Schrader R, Thornton JM (2009) Small Molecule Subgraph Detector (SMSD) toolkit. J Cheminform 1: 12. doi:10.1186/1758-2946-1-12.

108. Rogers DJ, Tanimoto TT (1960) A Computer Program for Classifying Plants. Science 132: 1115–1118. doi:10.1126/science.132.3434.1115.

109. O'Dwyer JP, Kembel SW, Green JL (2012) Phylogenetic diversity theory sheds light on the structure of microbial communities. PLoS Comput Biol 8: e1002832. doi:10.1371/journal.pcbi.1002832.

110. Mylona P, Pawlowski K, Bisseling T (1995) Symbiotic Nitrogen Fixation. 7: 869–885.

111. Vitousek PM, Cassman KEN, Cleveland C, Field CB, Grimm NB, et al. (2002) Towards an ecological understanding of biological nitrogen fixation. Biogeochemistry 57: 1–45. doi:10.1023/A:1015798428743.

112. Dodds WK, Gudder DA, Mollenhauer D (1995) The Ecology of Nostoc. J Phycol 31: 2–18. doi:10.1111/j.0022-3646.1995.00002.x.

113. Lavania M, Katoch K, Katoch VM, Gupta AK, Chauhan DS, et al. (2008) Detection of viable Mycobacterium leprae in soil samples: insights into possible sources of transmission of leprosy. Infect Genet Evol 8: 627–631. doi:10.1016/j.meegid.2008.05.007.

114. Young JS, Gormley E, Wellington EMH (2005) Molecular detection of Mycobacterium bovis and Mycobacterium bovis BCG (Pasteur) in soil. Appl Environ Microbiol 71: 1946–1952. doi:10.1128/AEM.71.4.1946-1952.2005.

115. Tyerman JG, Bertrand M, Spencer CC, Doebeli M (2008) Experimental demonstration of ecological character displacement. BMC Evol Biol 8: 34. doi:10.1186/1471-2148-8-34.

116. Herron MD, Doebeli M (2013) Parallel evolutionary dynamics of adaptive diversification in Escherichia coli. PLoS Biol 11: e1001490. doi:10.1371/journal.pbio.1001490.

117. Friesen ML, Saxer G, Travisano M, Doebeli M (2004) Experimental evidence for sympatric ecological diversification due to frequency-dependent competition in Escherichia coli. Evolution 58: 245–260.

118. Ahn MY, Shin KH, Kim DH, Jung EA, Toida T, et al. (1998) Characterization of a Bacteroides species from human intestine that degrades glycosaminoglycans. Can J Microbiol 44: 423–429.

119. Sava IG, Zhang F, Toma I, Theilacker C, Li B, et al. (2009) Novel interactions of glycosaminoglycans and bacterial glycolipids mediate binding of enterococci to human cells. J Biol Chem 284: 18194–18201. doi:10.1074/jbc.M901460200.

120. Hedrich S, Schlömann M, Johnson DB (2011) The iron-oxidizing proteobacteria. Microbiology 157: 1551–1564. doi:10.1099/mic.0.045344-0.

121. Archibald F (1983) Lactobacillus plantarum, an organism not requiring iron. FEMS Microbiol Lett 19: 29–32. doi:10.1111/j.1574-6968.1983.tb00504.x.

122. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, et al. (2013) Insights into the phylogeny and coding potential of microbial dark matter. Nature 499: 431–437. doi:10.1038/nature12352.

123. Manor O, Levy R, Borenstein E (2014) Mapping the Inner Workings of the Microbiome: Genomic- and Metagenomic-Based Study of Metabolism and Metabolic Interactions in the Human Microbiome. Cell Metab 20: 742–752. doi:10.1016/j.cmet.2014.07.021.

124. Nawrocki EP, Kolbe DL, Eddy SR (2009) Infernal 1.0: inference of RNA alignments. Bioinformatics 25: 1335–1337.

125. Cole JR, Wang Q, Cardenas E, Fish J, Chai B, et al. (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res 37: D141–D145.

126. Doncheva NT, Assenov Y, Domingues FS, Albrecht M (2012) Topological analysis and interactive visualization of biological networks and protein structures. Nat Protoc 7: 670–685. doi:10.1038/nprot.2012.004.

127. Kashtan N, Itzkovitz S, Milo R, Alon U (2004) Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. Bioinformatics 20: 1746–1758. doi:10.1093/bioinformatics/bth163.

128. Freeman L (1977) A set of measures of centrality based on betweenness. Sociometry 40: 35–41.

129. Cohen O, Ashkenazy H, Burstein D, Pupko T (2012) Uncovering the co-evolutionary network among prokaryotic genes. Bioinformatics 28: i389–i394. doi:10.1093/bioinformatics/bts396.

130. Wagner A (2009) Evolutionary constraints permeate large metabolic networks. BMC Evol Biol 9: 231. doi:10.1186/1471-2148-9-231.

131. Gordon JI, Klaenhammer TR (2011) A rendezvous with our microbes. Proc Natl Acad Sci U S A 108 Suppl : 4513–4515. doi:10.1073/pnas.1101958108.

132. Schink B, Stams AJM (2006) The Prokaryotes: 309–335. doi:10.1007/0-387-30742-7.

133. Morris J, Lenski R, Zinser E (2012) The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. MBio 3. doi:10.1128/mBio.00036-12.Updated.

134. Oliveira NM, Niehus R, Foster KR (2014) Evolutionary limits to cooperation in microbial communities. Proc Natl Acad Sci U S A: 201412673. doi:10.1073/pnas.1412673111.

135. Markowitz V, Chen I, Chu K (2012) IMG/M: the integrated metagenome data management and comparative analysis system. Nucleic Acids Res 40: 123–129. doi:10.1093/nar/gkr975.

136. Markowitz VM, Ivanova N, Palaniappan K, Szeto E, Korzeniewski F, et al. (2006) An experimental metagenome data management and analysis system. Bioinformatics 22: e359–e367. doi:10.1093/bioinformatics/btl217.

137. Laue H, Cook AM (2000) Biochemical and molecular characterization of taurine:pyruvate aminotransferase from the anaerobe Bilophila wadsworthia. Eur J Biochem 267: 6841–6848.

138. Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proc Natl Acad Sci 74: 5088–5090. doi:10.1073/pnas.74.11.5088.

139. Ramulu HG, Groussin M, Talla E, Planel R, Daubin V, et al. (2014) Ribosomal proteins: toward a next generation standard for prokaryotic systematics? Mol Phylogenet Evol 75: 103–117. doi:10.1016/j.ympev.2014.02.013.

140. Rajagopala S V, Titz B, Goll J, Parrish JR, Wohlbold K, et al. (2007) The protein network of bacterial motility. Mol Syst Biol 3: 128. doi:10.1038/msb4100166.

141. Giesy JP, Kannan K, Jones PD (2001) Global biomonitoring of perfluorinated organics. ScientificWorldJournal 1: 627–629. doi:10.1100/tsw.2001.342.

142. Holert J, Yücel O, Suvekbala V, Kulić Z, Möller H, et al. (2014) Evidence of distinct pathways for bacterial degradation of the steroid compound cholate suggests the potential for metabolic interactions by interspecies cross-feeding. Environ Microbiol 16: 1424–1440. doi:10.1111/1462-2920.12407.

143. Hugenholtz P (2002) Exploring prokaryotic diversity in the genomic era. Genome Biol 3: REVIEWS0003.

144. Manor O, Borenstein E (2015) MUSiCC: A marker genes based framework for metagenomic normalization and accurate profiling of gene abundances in the microbiome. Genome Biol In press.

145. GAUSE GF (1932) Experimental Studies on the Struggle for Existence: I. Mixed Population of Two Species of Yeast. J Exp Biol 9: 389–402.

146. Thiele I, Heinken A, Fleming R (2013) A systems biology approach to studying the role of microbes in human health. Curr Opin Biotechnol 24: 4–12. doi:10.1016/j.copbio.2012.10.001.

147. Heinken A, Sahoo S, Fleming R, Thiele I (2013) Systems-level characterization of a host-microbe metabolic symbiosis in the mammalian gut. Gut Microbes 4: 1–13.

148. Segrè D, Vitkup D, Church GM (2002) Analysis of optimality in natural and perturbed metabolic networks. Proc Natl Acad Sci U S A 99: 15112–15117. doi:10.1073/pnas.232349399.

149. Shlomi T, Berkman O, Ruppin E (2005) Regulatory on/off minimization of metabolic flux changes after genetic perturbations. Proc Natl Acad Sci U S A 102: 7695–7700. doi:10.1073/pnas.0406346102.

150. Ishii N, Nakahigashi K, Baba T, Robert M, Soga T, et al. (2007) Multiple high-throughput analyses monitor the response of E. coli to perturbations. Science 316: 593–597. doi:10.1126/science.1132067.

151. Simberloff D, Wilson E (1969) Experimental zoogeography of islands: the colonization of empty islands. Ecology 50: 278–296.

152. Chalmers NI, Palmer RJ, Cisar JO, Kolenbrander PE (2008) Characterization of a Streptococcus sp.-Veillonella sp. community micromanipulated from dental plaque. J Bacteriol 190: 8145–8154. doi:10.1128/JB.00983-08.

153. Periasamy S, Chalmers NI, Du-Thumm L, Kolenbrander PE (2009) Fusobacterium nucleatum ATCC 10953 requires Actinomyces naeslundii ATCC 43146 for growth on saliva in a three-species community that includes Streptococcus oralis 34. Appl Environ Microbiol 75: 3250–3257. doi:10.1128/AEM.02901-08.

154. Periasamy S, Kolenbrander PE (2009) Aggregatibacter actinomycetemcomitans builds mutualistic biofilm communities with Fusobacterium nucleatum and Veillonella species in saliva. Infect Immun 77: 3542–3551. doi:10.1128/IAI.00345-09.

155. Periasamy S, Kolenbrander PE (2010) Central role of the early colonizer Veillonella sp. in establishing multispecies biofilm communities with initial, middle, and late colonizers of enamel. J Bacteriol 192: 2965–2972. doi:10.1128/JB.01631-09.

156. Chao A, Chazdon RL, Colwell RK, Shen T-J (2005) A new statistical approach for assessing similarity of species composition with incidence and abundance data. Ecol Lett 8: 148–159. doi:10.1111/j.1461-0248.2004.00707.x.

157. Bray JR, Curtis JT (1957) An ordination of the upland forest communities of southern Wisconsin. Ecol Monogr 27: 325–349. doi:10.2307/1942268.

158. Horn HS (1966) Measurement of “Overlap” in Comparative Ecological Studies. Am Nat 100: 419–424. doi:10.1086/282436.

159. Aitchison J (1982) The statistical analysis of compositional data. J R Stat Soc Ser B 44: 139–177.

160. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, et al. (2011) Enterotypes of the human gut microbiome. Nature 473: 174–180. doi:10.1038/nature09944.

161. Kreimer A, Doron-Faigenboim A, Borenstein E, Freilich S (2012) NetCmpt: a network-based tool for calculating the metabolic competition between bacterial species. Bioinformatics 28: 2195–2197. doi:10.1093/bioinformatics/bts323.

162. Ebenhöh O, Handorf T, Heinrich R (2004) Structural analysis of expanding metabolic networks. Genome Inform 15: 35–45.