Pathway-based and phylogenetically adjusted quantification of metabolic interaction between microbial species

Tony J. Lam[1], Moses Stamboulian[1], Wontack Han[1], Yuzhen Ye[1*],

1 Luddy School of Informatics, Computing and Engineering, Bloomington, IN, USA

* Corresponding Author: yye@indiana.edu

Abstract

Microbial community members exhibit various forms of interactions. Taking advantage of the increasing availability of microbiome data, many computational approaches have been developed to infer bacterial interactions from the co-occurrence of microbes across diverse microbial communities. Additionally, the introduction of genome-scale metabolic models have also enabled the inference of metabolic interactions, such as competition and cooperation, between bacterial species. By nature, phylogenetically similar microbial species are likely to share common functional profiles or biological pathways due to their genomics similarity. Without properly factoring out the phylogenetic relationship, any estimation of the competition and cooperation based on functional/pathway profiles may bias downstream applications.

To address these challenges, we developed a novel approach for estimating the competition and complementarity indices for a pair of microbial species, adjusted by their phylogenetic distance. An automated pipeline, PhyloMint, was implemented to construct competition and complementarity indices from genome scale metabolic models derived from microbial genomes. Application of our pipeline to 2,815 human-gut bacteria showed high correlation between phylogenetic distance and metabolic competition/cooperation indices among bacteria. Using a discretization approach, we were able to detect pairs of bacterial species with cooperation scores significantly higher than the average pairs of bacterial species with similar phylogenetic distances. A network community analysis of high metabolic cooperation but low competition reveals distinct modules of bacterial interactions. Our results suggest that niche differentiation plays a dominant role in microbial interactions, while habitat filtering also plays a role among certain clades of bacterial species.

Author summary

Microbial communities, also known as microbiomes, are formed through the interactions of various microbial species. Utilizing genomic sequencing, it is possible to infer the compositional make-up of communities as well as predict their metabolic interactions. However, because some species are more similarly related to each other, while others are more distantly related, one cannot directly compare metabolic relationships without first accounting for their phylogenetic relatedness. Here we developed a computational pipeline which predicts complimentary and competitive metabolic relationships between bacterial species, while normalizing for their phylogenetic relatedness. Our results show that phylogenetic distances are correlated with metabolic interactions, and factoring out such relationships can help better understand microbial interactions which drive community formation.

Introduction

Recent advances in microbiome research have accelerated the study of the composition and function of microbial communities associated with different environments and hosts. Studies have shown the association of microbial communities with human health and diseases including type 2 diabetes (1), and efficacy of treatment including immunotherapy to cancers (2). To reveal the mechanisms behind the microbiome-host interactions, it is important to understand how microbial species form communities and how the microbial communities interact with the host to mediate various biological processes (3).

Studying the principles underlying the structure and composition of microbial communities is of long-standing interest to microbial ecologists. The dynamics which govern microbial community assembly have been extensively debated, and it is disputed upon as to what extent the role of neutral or deterministic dynamics plays in microbial interactions (4; 5). Some studies support the neutral hypothesis, which assumes that community structure is determined by random processes (6). Other theories suggest that community assembly dynamics are govern by deterministic processes such as habitat filtering and niche differentiation (7; 8). While many studies focus on species abundances for studying community assembly, Bruke et al. (9) showed that the key level at which to address the community assembly may not be species, but rather the functional level of genes. Both niche and neutral processes are likely to affect the assembly of complex microbial communities.

Some studies have shown that microbial communities tend to be more phylogenetically clustered than expected by chance, harboring groups of closely related taxa that exhibit microscale differences in genomic diversity (10; 11; 12). One study of marine bacterial communities at various locations reported that local communities are phylogenetically different from each other and they tend to be phylogenetically clustered (12). However, some microbial communities have also shown the opposite patterns, in which taxa are less clustered and are less related than expected by chance (13; 14). Together, these studies have explored the relationship between functional distances/metabolic overlap with phylogenetic relatedness, and they have given rise to competing theories of 'habitat-filtering' and 'niche differentiation': habitat filtering suggests that dominant species exhibit similar functional traits, whereas niche differentiation says that phylogenetically similar species are unable to co-exist due to similar traits and resource overlap (3). Nevertheless, methods have been developed for inference of bacterial interaction network based on the assumption that phylogenetically related species tend to co-exists. For example, Lo et al. (15) developed phylogenetic graphical lasso approach for bacterial community detection, based on the assumption that phylogenetically correlated microbial species are more likely to interact to each other.

The study of microbial interactions and the dynamics which govern such interactions are important in providing insights to community assembly and ultimately processes which influence host health and disease. Insights into community cooperation and competition may also uncover symbiotic and antagonistic relationships and can be used to provide prospective candidates for probiotics. Leveraging the increasing availability of microbiome datasets, novel statistical and computational methods have been developed to infer bacterial interaction networks from co-occurrence information. Some examples include, SparCC (16), a tool to infer correlations by correcting for compositional data. Conner et al. demonstrated the importance of using null model to infer microbial co-occurrence networks (17). Mandakovic and colleagues compared microbial co-occurrence networks representing bacterial soil communities from different environments to determine the impact of a shift in environmental variables on the community's taxonomic composition and their relationships (18). MDiNE is another recently developed model for estimating differential co-occurrence networks in microbiome studies (19). Notably, Faust et al. (20) applied generalized boosted linear models to infer thousands of significant co-occurrence and co-exclusion relationships between 197 clades occurring throughout the human microbiomes; their study revealed reverse correlation between functional similarity and phylogenetic distance among bacterial species, which is unsurprising. Despite of the numerous advances, it has been considered difficult to infer microbial community structure based on co-occurrence network approaches (21).

Functional profiles or biological pathways inferred from genomic sequences of the microbial species can provide mechanistic information about the functional traits of the microbes and potential cross-feeding. Genome-scale metabolic models (GEMS) potentially could provide mechanistic explanations to the association of bacterial species that are discovered by analyzing their co-occurrence in diverse microbial communities (22). Many automated tools (23; 24; 25; 26) have been made available for genome scale metabolic reconstructions (GENREs), however to get quality models these automated methods often require additional manual refinement including checks for stoichiometric consistency, defined media, and gap filling (27). The challenges of manual curation often make it difficult to construct GEMs for a large consortium of microbes. Notably, Machado et al. (28) developed an automated tool called CarveMe, which uses a top-down approach to build species and community level metabolic models which the authors claim is able to produce comparable results to other tools while also reducing manual intervention (28; 29). The ability to predict metabolic network of microbial members through GENREs has led some studies to focus on inferring levels and types of interaction among microbial species via metabolic models. Levy and Borenstein (30) introduced pairwise indices of metabolic interaction: the metabolic competition index and complementarity index, which are computed based on the overlapping and complementarity of the compounds that are contained in the metabolic models, respectively. By analyzing the metabolic interactions among 154 human-associated bacterial species and comparing the computed indices with observed species co-occurrence in microbiomes, the authors concluded that species tend to co-occur across individuals more frequently with species with which they strongly compete, suggesting that microbial assembly is dominated by habitat filtering (30). Similar metrics have been introduced to quantify the metabolic cooperation and competition between bacterial species, such as MIP (metabolic interaction potential) and MRO (metabolic resource overlap) (22).

By nature, two phylogenetically-close microbial species share similar functional profiles or biological pathways due to their genomic similarity. Additionally, co-evolutionary studies have also shown that comparative analyses between species cannot be assumed to be statistically independent, as comparative data of similarly related species correlate with each other due to shared ancestry (31; 32; 33). Thus, without factoring out the phylogenetic relationship (the confounding factor), any estimation of the competition and cooperation tendency based on function/pathway profiles may be biased and cause problems in downstream applications. In this study, we focused on the large collection of human gut-related genomes (including reference genomes and genomes assembled from metagenomic sequences, MAGs). We implemented an automated pipeline (called PhyloMInt) for genome scale pathway reconstruction and for computing competition and cooperation scores based on the reconstructed pathways. Our results showed correlation between phylogenetic distance and metabolic competition/cooperation indices, indicating the importance of normalizing these indices by the phylogenetic distance between underlying microbial species. Using a discretization approach, we were able to detect pairs of bacterial species with cooperation scores significantly higher than the average pairs of bacterial species with similar phylogenetic distances. We further built a network of human-gut microbes based on cooperation and competition indices, and we discuss some of the results we derived by analyzing the network.

Results

Evaluation of the performance of GENREs on incomplete genomes

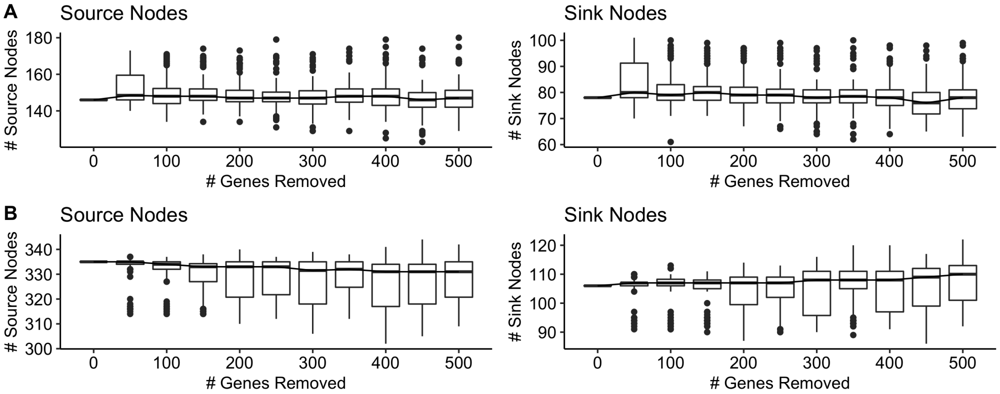

Fig 1. Number of source and sink nodes in metabolic networks inferred from genome scale metabolic reconstructions with randomly removed genes. (A) GENRE of *Mycobacterium tuberculosis* H37Rv (Accession: PRJNA57777). (B) GENRE of *Escherichia coli* str. K-12 (Accession: SAMN02604091).

We first evaluated the effect of using MAGs for genome-scale metabolic reconstructions when genomes are incomplete. We tested the robustness of GENRE using simulated incomplete genomes by removing genes from complete genomes and evaluating the resulting GENREs. We simulated incomplete genomes with 50, 100, $\cdots$, 500 genes randomly removed from each genome, respectively, and for each setting we repeated 100 simulations. Using the reconstructed GENREs, we were able to analyze the distribution of source and sink nodes (the calculation of the complementarity and competition indices is dependent on the identification of source and sink nodes in the metabolic model) within the metabolic networks reconstructed from incomplete genomes. Our empirical analysis shows the mean source and sink metabolites in GENREs remained relatively stable in respect to the removal of genes (details can be found in supplementary data). This provided us confidence that the incompleteness of the near-complete MAGs should have minimal impact on the calculation of the metabolic complementarity/competition indices.

Impact of phylogenetic relationship on microbial complementarity and competition indices

We applied our pipeline to analyze 2,815 human gut related MAGs and computed their pairwise competition and complementarity scores (about 8M directed pairs). As shown in Figure 2A, we see a positive relationship between the metabolic complementarity of bacterial species and their phylogenetic distances. In contrast, we see in Figure 2B there is a negative relationship between metabolic competition of bacterial species and phylogenetic distance. Our results are consistent with other previous studies of functional and metabolic relationships with phylogenetic distances (20; 22; 34). And they support the theory of niche differentiation, which states that phylogenetically close species are more likely to compete with each other due to their shared traits and resource overlap, leading to less probability of their co-existence.

Due to the non-zero correlation between metabolic interactions and phylogenetic distances, comparing complementarity and competition between species pairs without accounting for their phylogenetic relationships confounds such comparisons. Here we

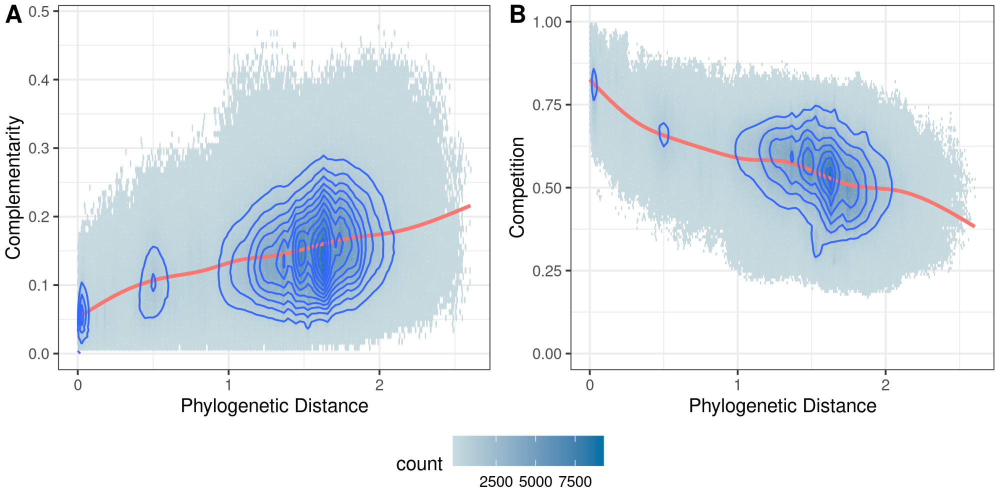

Fig 2. Hexagonal binned plots of (A) metabolic complementarity index and (B) metabolic cooperation index, versus phylogenetic distance with density contours. The plots were fitted with a generalized additive model (red line).

demonstrate a discretization approach for the identification of statistically significant complementary species pairs as a method for accounting/correcting for phylogenetic distances. To discretize comparisons across continuous phylogenetic distances, pairwise indices were binned by their phylogenetic distances. Outliers are then identified within each bin, which are likely pairs of bacteria with statistically significant complementary or competitive interactions.

Identification of potentially collaborative or competing pairs of gut bacteria from metabolic outliers

To explore the relationship between complementary and competitive pairs, we compared their respective Z-scores (Figure 3). Significant outliers were selected using a Z-score threshold of ±2.698 as proposed by Tukey (35). A total of 60,116 directed pairs were identified as positive complementary outliers. Additionally, 7,769 and 44,409 competitive positive and negative directed pairs of outliers were identified, respectively. Unsurprisingly, most pairs were centered around a Z-score of zero and no pairs were simultaneously significant for both complementarity and competition, simultaneously.

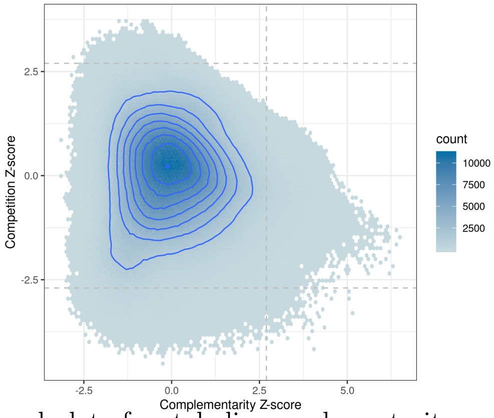

Fig 3. Hexagonal binned plot of metabolic complementarity and competition Z-scores with density contours.

We analyzed bacteria pairs belonging to the same genus or family that have significantly high complementarity scores to better understand how taxonomic similarity correlates with metabolic cooperation. At the genus level, 140,152 directed pairs were identified; and at the family level, 233,555 directed pairs were identified. Of the pairs belonging to the same genus or family, 1,230 and 5,190 were identified as significant complementary outliers, respectively. These taxonomically similar bacteria pairs have the potential to cooperate in gut microbiomes (see detailed lists of species pairs in the supplementary data). The rarity of significant outliers with the same taxonomic classifications suggests that for most taxonomically similar pairs at the genus and family level, niche differentiation plays an integral role in community assembly.

Exploration of microbial complementarity/competition network: community structure and composition

To explore community assembly dynamics, we constructed a directed graph of bacterial species where bacteria are the nodes and a directed edge is added between two bacteria if they have a high metabolic complementarity (Z-score > 2.698) and low metabolic competition (Z-score < -1.000); here we relaxed the Z-score of competition indicies to -1.000 in-order to focus our analysis towards species pairs with greater complementarity while still constraining the analysis to a degree of low competition observed between species pairs. Using Infomap (36) to analyze the network, we were able to identify two main community modules (Figure 4). The larger community module (shown on the right in Figure 4) was populated with many multi-layer sub-modules, which featured majority of the significantly cooperating bacteria. Interestingly the smaller community module (shown on the left in Figure 4) exclusively contained *Bifidobacterium spp.* (e.g. *B. longum*, *B. bifidum*, *B. infantis*), suggesting that various *Bifidobacterium* species are metabolically complementary to each other, more-so than other phylogenetically similar taxa. Furthermore, a small fraction of community sub-modules within the right larger community module were also dominated by taxonomically similar genomes (i.e. *Helicobacter*, *Collinsella*, *Lachnospiraceae*, and *Ruminococcus*). We note that if complementarity scores were analyzed without correcting for phylogenetic distances, these significant complementarity scores of taxonomically similar bacteria would not be considered significant, thus emphasizing the importance of correcting for phylogenetic distances. The pattern of taxonomically related genomes forming community module is suggestive of habitat filtering characteristics within certain distinctive bacterial taxa (the details of the membership of the modules can be found in the supplementary data).

To further explore this, we analyzed the proportion of significantly cooperative bacteria with the same genus annotations. Our results show that more than half (42/76) of the taxa with 50 or more members within the same genus contained a significant number of metabolically complementary pairs; within genus proportion of taxa with significant pairs ranged from 0.02% to 15.9% (details of the specie pairs are available in the supplementary data). Together, these results show that while niche differentiation dominates a majority of metabolic interactions, we observe habitat filtering characteristics within certain bacterial taxa.

Discussion

Here we demonstrate a novel approach to identifying significant metabolic cooperators and competitors between bacterial species pairs. This approach builds upon previously developed metrics of metabolic complementation and cooperation (30; 37; 38) by identifying outlier pairs relative to their phylogenetic distances. As pairwise metabolic interactions are correlated with phylogenetic distance, it remains imperative to take into consideration their phylogenetic distances when making comparisons across different phylogenetic distances as such comparisons may confound comparisons.

Our analysis shows that metabolic cooperation exhibits a positive relationship with phylogenetic distance, whereas metabolic competition exhibits a negative relationship. These findings support the results from previous work that studied the relationship between phylogenetic relatedness and gene content, functional distance, and metabolic

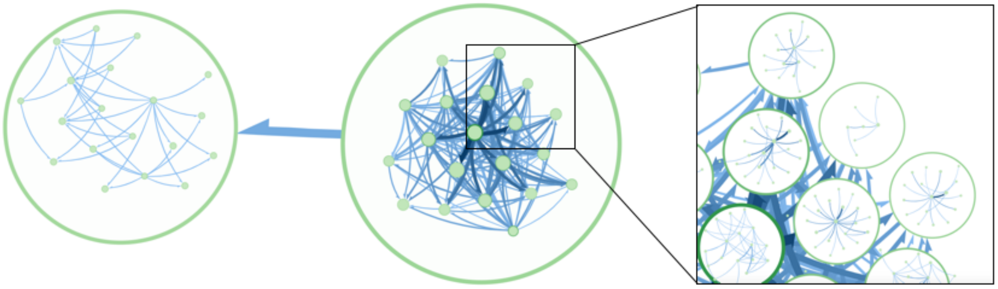

Fig 4. Community modules of significant complementarity outliers that exhibit low metabolic competition identified from human-gut related MAGs. Circular nodes represent predicted community modules and sub-modules of cooperative bacterial communities.

interactions (20; 22; 34). Together these observed relationships seem to support the theory of niche differentiation, where functional overlap discourages phylogenetically related species from co-existing. However, by taking into consideration the phylogenetic distance between pairs to identify metabolic outliers, we were able to identify significant intra-genus cooperation in several distinct taxa. The intra-genus modules may suggest that while most bacteria interactions display niche differentiation characteristics, some taxa exhibit habitat filtering. Notably, *Bifidobacterium* species were shown to form distinct community modules which suggest significant intra-genus cooperation compared to other taxa. These results support recent findings that suggest strains of *Bifidobacterium spp.* in infants have different nutrient profiles to support colonization of other specific *Bifidobacterium* species (39). The observation of both habitat filtering and niche differentiation characteristics suggests that in some cases both contribute to the dynamics of community assembly.

We note a few limitations of our approach. First, metabolic complementarity and competition indices are dependent on a given metabolic model. Completeness of GENREs are dependent on a variety of variables (e.g. the reconstruction tool and the genome completeness) that can have a significant impact on predicted metabolic interactions. Second, seed sets used to calculate the metabolic interaction indices do not represent required metabolites for growth, but rather represent a baseline of metabolites that in theory enable a given bacterium to produce any metabolite in their predicted metabolic network. As such, seed sets may influence the overestimation or underestimation of metabolic interactions between bacterial species. However, by integrating phylogenetic distances to normalize metabolic interaction indices we believe that our approach provides a more accurate prediction of metabolic interactions in comparison to other similar methods. Additionally, low abundant microbial species within microbiomes are not always well represented within metagenomic samples but may play key roles within a metabolic network. While we acknowledge that validation of this method remains difficult, the non-independent nature of comparative metricies between organisms due to shared ancestry provides a logical explanation as to the necessity to account for such confounding effects.

By decoupling phylogenetic distances between complementarity and competition indicies, we provide a method to explore statistically significant cooperating/competing species pairs within microbbiomes to better understand community assembly dynamics. Additionally, competition networks can be used to identify highly competitive species pairs, which may be useful for suggesting beneficial probiotic candidates. A future research direction is to integrate phylogenetically-corrected cooperation and competition scores with co-occurrence information to better address the challenges of identifying bacterial interactions through mechanistic insight.

Materials and Methods

Genome sequences of human-gut bacteria

To assemble the human-gut associated reference genomes, we collected genomes from two recent studies (40; 41). Bacterial genomes reported in (41) were compiled from two sources: a total of 617 genomes obtained from the human microbiome project (HMP) (42), and 737 whole genome-sequenced bacterial isolates, representing the Human Gastrointestinal Bacteria Culture Collection (HBC). These 737 bacterial genomes were assembled by culturing and purifying bacterial isolates of 20 fecal samples originating from different individuals (41). The bacterial genomes reported in (40) were generated and classified from a total of 92,143 metagenome assembled genomes (MAGs), among which a total of 1,952 binned genomes were characterized as non-overlapping with bacterial genomes reported. We were able to retrieve 612 out of 617 RefSeq sequences using the reported RefSeq IDs. We only included genomes with $> 80\%$ completeness and $< 5\%$ contamination (via CheckM (43)). Our final dataset for this study contains a total of 2,815 genomes/MAGs. Taxonomic annotation of these genomes/MAGs was done using GTDB-toolkit's least common ancestors approach (44).

Genome scale metabolic network reconstructions and analysis

Genome-scale metabolic network reconstructions (GENREs) for all genomes were constructed using CarveMe(28) with default parameters. Coding sequences (CDSs) of all input genomes were generated using FragGeneScan(45) to be used as input for CarveMe. Briefly, CarveMe is a genome-scale metabolic model reconstruction tool which utilizes a universal model for a top-down approach to build GENREs. In contrast to conventional bottom-up methods which require well defined growth media, manual curation and gap-filling, the top down approach of CarveMe removes reactions and metabolites inferred to be not present in the manually curated universal template.

Phylogenetic distance

To compute pairwise evolutionary distances between gut bacteria, we first inferred a phylogeny covering all participating genomes using FastTree(46). A total of 120 bacterial marker genes were used to infer these phylogeny. The 120 marker genes used are ubiquitous among bacterial species and are shown to occur as single copies and less susceptible to horizontal gene transfer(47). Amino acid sequence of protein coding genes were searched using HMMER3(48) against a 120 HMM model database of marker genes received from Pfam(49) and TIGRfam databases(50). Sequences extracted from each HMM model were individually aligned using hmmalign, which were later concatenated to form the final alignment. Poorly aligned regions were removed from the concatenated alignment and a final phylogeny was inferred using FastTree under WAG + GAMMA models.

Species interaction indexes

To estimate potential metabolic cooperation and competition between bacterial species, we need to know their nutritional profiles, which however are unavailable for most of the gut bacteria. Similar to the approach reported in (30; 51), we use the compound *seed* set of each species as a proxy for its nutritional profile: the seed set of a metabolic network is defined as the minimal subset of the compounds that cannot be synthesized from other compounds in the network (due to lack of the corresponding enzymes, and hence are exogenously acquired) but their existence permits the production of all other compounds in the network.

We implemented a pipeline for computing metabolic interaction indices from genome sequences. Our pipeline uses 1) CarveMe for building genome-scale metabolic models from genome sequences, b) NetworkX (52) to identity seed compounds, and c) our own implementation (in Python) of the approaches for computing metabolic competition and complimentary indices given two genome-scale metabolic models. We call our pipeline PhyloMInt (Phylogenetically-adjusted Metabolic Interaction indices).

Seed set identification

Utilizing NetworkX v2.2 (52), strongly connected components (SCC) within the GENREs are identified. Confidence levels are assigned for all compounds relative to their SCC size, where the confidence level (C) is denoted as:

$$C = \frac{1}{(Component\ Size)} \tag{1}$$

The confidence level is representative of the confidence that a given compound belongs to the seed set. A threshold of $C \geq 0.2$ was used to select compounds to be regarded as compounds part of a given 'seed set' of a given organism as specified by (51).

Metabolic competition and complementarity indices

Given two genome-scale metabolic models (GEMs) A and B, their Metabolic Competition Index ($MI_{Competition}$) is calculated as the fraction of A's seed set that is also in B's seed set, normalized by the weighted sum of the confidence score (30; 38). $MI_{Competition}$ estimates the baseline metabolic overlap between two given metabolic networks.

$$MI_{Competition} = \frac{\sum C(SeedSet_A \cap SeedSet_B)}{\sum C(SeedSet_A)} \tag{2}$$

Complementarity Index ($MI_{Complementarity}$) is calculated as the fraction of A's seed set that is found within B's metabolic network but not part of B's seed set, normalized by the number of A's seed set in B's entire metabolic network (30; 37). $MI_{Complementarity}$ represents the potential for A's to utilize the potential metabolic output of B.

$$MI_{Complementarity} = \frac{|SeedSet_A \cap \neg SeedSet_B|}{|SeedSet_A \cap (SeedSet_B \cup \neg SeedSet_B)|} \tag{3}$$

We note that the competition and complementarity indices are asymmetric.

Phylogenetic normalization and outlier detection

Pairwise metabolic complementarity and competition indices between species pairs are plotted against their predicted phylogenetic distance. While methods of outlier detection for continuous data exists, local peaks and troughs of indices relative to phylogenetic distance make it difficult to identify local outliers. Thus, we utilize a binning approach to limit outlier detection to localized values. Both metabolic complementarity and competition indexes use a two-step binning process to bin pairwise observations, first by using a fixed phylogenetic distance interval of 0.01, followed by merging bins which are smaller than a prespecified size. Here we used the first bin size as the reference. Bins were merged with the closest preceding bin satisfying our minimum bin size threshold. To identify metabolic complementarity and cooperation outliers within each phylogenetic distance bin, we calculate the Z-score within each bin respectively. Tukey's method for outlier detection (equivalent to a Z-score threshold ±2.698) (35) was utilized to identify significant outliers.

Network construction and community detection

To build a metabolic complementarity/competition network, species pairs are represented as nodes within the network. Identified significant outliers were used to construct a network of gut bacteria, in which for any pair of species A and B, a directed edge is added between A and B (from A to B), if A and B have significantly high complementarity score but low competition score. Using the adjacency list of the directed graph, a local installation of Infomap(36) (with the parameters: –directed –zero-based-numbering –num-trials 10) was utilized to identify community interaction modules within our dataset. Infomap is a random walk based approach for community detection, and it provides a user friendly interface for visualization and exploration of the network and community structure (https://www.mapequation.org/navigator).

Data and software availability

Implementation and data are available at https://github.com/mgtools/PhyloMint.

References

1. Zhao L, Zhang F, Ding X, et al. Gut bacteria selectively promoted by dietary fibers alleviate type 2 diabetes. Science. 2018;359(6380):1151–1156.
2. Routy B, Le Chatelier E, et al. Gut microbiome influences efficacy of PD-1–based immunotherapy against epithelial tumors. Science. 2018;359(6371):91–97.
3. Nemergut DR, Schmidt SK, Fukami T, O'Neill SP, Bilinski TM, Stanish LF, et al. Patterns and processes of microbial community assembly. Microbiol Mol Biol Rev. 2013;77(3):342–356.
4. Zhou J, Ning D. Stochastic community assembly: does it matter in microbial ecology? Microbiol Mol Biol Rev. 2017;81(4):e00002–17.
5. Powell JR, Karunaratne S, Campbell CD, Yao H, Robinson L, Singh BK. Deterministic processes vary during community assembly for ecologically dissimilar taxa. Nature communications. 2015;6(1):1–10.
6. Woodcock S, Van Der Gast CJ, Bell T, Lunn M, Curtis TP, Head IM, et al. Neutral assembly of bacterial communities. FEMS microbiology ecology. 2007;62(2):171–180.
7. Dumbrell AJ, Nelson M, Helgason T, Dytham C, Fitter AH. Relative roles of niche and neutral processes in structuring a soil microbial community. The ISME journal. 2010;4(3):337–345.
8. Wong HL, Smith DL, Visscher PT, Burns BP. Niche differentiation of bacterial communities at a millimeter scale in Shark Bay microbial mats. Scientific reports. 2015;5:15607.
9. Burke C, Steinberg P, Rusch D, Kjelleberg S, Thomas T. Bacterial community assembly based on functional genes rather than species. Proceedings of the National Academy of Sciences. 2011;108(34):14288–14293.
10. Horner-Devine MC, Bohannan BJ. Phylogenetic clustering and overdispersion in bacterial communities. Ecology. 2006;87(sp7):S100–S108.
11. Bryant JA, Lamanna C, Morlon H, Kerkhoff AJ, Enquist BJ, Green JL. Microbes on mountainsides: contrasting elevational patterns of bacterial and plant diversity. Proceedings of the National Academy of Sciences. 2008;105(Supplement 1):11505–11511.
12. Pontarp M, Canbäck B, Tunlid A, Lundberg P. Phylogenetic analysis suggests that habitat filtering is structuring marine bacterial communities across the globe. Microbial ecology. 2012;64(1):8–17.
13. Thompson JR, Pacocha S, Pharino C, Klepac-Ceraj V, Hunt DE, Benoit J, et al. Genotypic diversity within a natural coastal bacterioplankton population. Science. 2005;307(5713):1311–1313.
14. Chaffron S, Rehrauer H, Pernthaler J, Von Mering C. A global network of coexisting microbes from environmental and whole-genome sequence data. Genome research. 2010;20(7):947–959.
15. Lo C, Marculescu R. PGLasso: Microbial Community Detection through Phylogenetic Graphical Lasso. arXiv preprint arXiv:180708039. 2018;.
16. Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. PLoS computational biology. 2012;8(9).
17. Connor N, Barberán A, Clauset A. Using null models to infer microbial co-occurrence networks. PloS one. 2017;12(5).

18. Mandakovic D, Rojas C, Maldonado J, Latorre M, Travisany D, Delage E, et al. Structure and co-occurrence patterns in microbial communities under acute environmental stress reveal ecological factors fostering resilience. Scientific reports. 2018;8(1):5875.
19. McGregor K, Labbe A, Greenwood CMT. MDiNE: a model to estimate differential co-occurrence networks in microbiome studies. Bioinformatics. 2019;doi:10.1093/bioinformatics/btz824.
20. Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, Raes J, et al. Microbial co-occurrence relationships in the human microbiome. PLoS computational biology. 2012;8(7):e1002606.
21. Hirano H, Takemoto K. Difficulty in inferring microbial community structure based on co-occurrence network approaches. BMC bioinformatics. 2019;20(1):329.
22. Zelezniak A, Andrejev S, Ponomarova O, Mende DR, Bork P, Patil KR. Metabolic dependencies drive species co-occurrence in diverse microbial communities. Proceedings of the National Academy of Sciences. 2015;112(20):6449–6454.
23. Devoid S, Overbeek R, DeJongh M, Vonstein V, Best AA, Henry C. Automated genome annotation and metabolic model reconstruction in the SEED and Model SEED. In: Systems Metabolic Engineering. Springer; 2013. p. 17–45.
24. Agren R, Liu L, Shoaie S, Vongsangnak W, Nookaew I, Nielsen J. The RAVEN toolbox and its use for generating a genome-scale metabolic model for Penicillium chrysogenum. PLoS computational biology. 2013;9(3).
25. Dias O, Rocha M, Ferreira EC, Rocha I. Reconstructing genome-scale metabolic models with merlin. Nucleic acids research. 2015;43(8):3899–3910.
26. Karp PD, Latendresse M, et al. Pathway Tools version 19.0 update: software for pathway/genome informatics and systems biology. Briefings in bioinformatics. 2016;17(5):877–890.
27. Thiele I, Palsson BØ. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nature protocols. 2010;5(1):93.
28. Machado D, Andrejev S, Tramontano M, Patil KR. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. Nucleic Acids Research. 2018;46(15):7542–7553. doi:10.1093/nar/gky537.
29. Mendoza SN, Olivier BG, Molenaar D, Teusink B. A systematic assessment of current genome-scale metabolic reconstruction tools. Genome Biology. 2019;20(1). doi:10.1186/s13059-019-1769-1.
30. Levy R, Borenstein E. Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules. Proceedings of the National Academy of Sciences. 2013;110(31):12804–12809. doi:10.1073/pnas.1300926110.
31. Dutheil JY. Detecting coevolving positions in a molecule: why and how to account for phylogeny. Briefings in Bioinformatics. 2011;13(2):228–243. doi:10.1093/bib/bbr048.
32. Rezende EL, Diniz-Filho JAF. Phylogenetic analyses: comparing species to infer adaptations and physiological mechanisms. Comprehensive Physiology. 2011;2(1):639–674.
33. Cope AL, O'Meara B, Gilchrist MA. Gene Expression of Functionally-Related Genes Coevolves Across Fungal Species: Detecting Coevolution of Gene Expression Using Phylogenetic Comparative Methods. 2019;doi:10.1101/844472.
34. Hester ER, Jetten MS, Welte CU, Lücker S. Metabolic overlap in environmentally diverse microbial communities. Frontiers in genetics. 2019;10:989.
35. Tukey JW. Exploratory Data Analysis. Pearson; 1977.

36. Rosvall M, Bergstrom CT. Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences. 2008;105(4):1118–1123. doi:10.1073/pnas.0706851105.

37. Levy R, Carr R, Kreimer A, Freilich S, Borenstein E. NetCooperate: a network-based tool for inferring host-microbe and microbe-microbe cooperation. BMC bioinformatics. 2015;16(1):164.

38. Kreimer A, Doron-Faigenboim A, Borenstein E, Freilich S. NetCmpt: a network-based tool for calculating the metabolic competition between bacterial species. Bioinformatics. 2012;28(16):2195–2197.

39. Vatanen T, Plichta DR, Somani J, Münch PC, Arthur TD, Hall AB, et al. Genomic variation and strain-specific functional adaptation in the human gut microbiome during early life. Nature microbiology. 2019;4(3):470–479.

40. Almeida A, Mitchell AL, Boland M, Forster SC, Gloor GB, Tarkowska A, et al. A new genomic blueprint of the human gut microbiota. Nature. 2019;568(7753):499.

41. Forster SC, Kumar N, Anonye BO, Almeida A, Viciani E, Stares MD, et al. A human gut bacterial genome and culture collection for improved metagenomic analyses. Nature biotechnology. 2019;37(2):186.

42. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The human microbiome project. Nature. 2007;449(7164):804.

43. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome research. 2015;25(7):1043–1055.

44. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil PA, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nature biotechnology. 2018;.

45. Rho M, Tang H, Ye Y. FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Research. 2010;38(20):e191–e191. doi:10.1093/nar/gkq747.

46. Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. PloS one. 2010;5(3):e9490.

47. Parks DH, Rinke C, Chuvochina M, Chaumeil PA, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nature microbiology. 2017;2(11):1533–1542.

48. Finn RD, Clements J, Arndt W, Miller BL, Wheeler TJ, Schreiber F, et al. HMMER web server: 2015 update. Nucleic acids research. 2015;43(W1):W30–W38.

49. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic acids research. 2014;42(D1):D222–D230.

50. Haft DH, Selengut JD, White O. The TIGRFAMs database of protein families. Nucleic acids research. 2003;31(1):371–373.

51. Borenstein E, Kupiec M, Feldman MW, Ruppin E. Large-scale reconstruction and phylogenetic analysis of metabolic environments. Proceedings of the National Academy of Sciences. 2008;105(38):14482–14487. doi:10.1073/pnas.0806162105.

52. Hagberg A, Swart P, S Chult D. Exploring network structure, dynamics, and function using NetworkX. Los Alamos National Lab.(LANL), Los Alamos, NM (United States); 2008.