scientific reports

Check for updates

OPEN

Meta-analysis of microbiome association networks reveal patterns of dysbiosis in diseased microbiomes

Tony J. Lam & Yuzhen Ye✉

The human gut microbiome is composed of a diverse and dynamic population of microbial species which play key roles in modulating host health and physiology. While individual microbial species have been found to be associated with certain disease states, increasing evidence suggests that higher-order microbial interactions may have an equal or greater contribution to host fitness. To better understand microbial community dynamics, we utilize networks to study interactions through a meta-analysis of microbial association networks between healthy and disease gut microbiomes. Taking advantage of the large number of metagenomes derived from healthy individuals and patients with various diseases, together with recent advances in network inference that can deal with sparse compositional data, we inferred microbial association networks based on co-occurrence of gut microbial species and made the networks publicly available as a resource (GitHub repository named GutNet). Through our meta-analysis of inferred networks, we were able to identify network-associated features that help stratify between healthy and disease states such as the differentiation of various bacterial phyla and enrichment of Proteobacteria interactions in diseased networks. Additionally, our findings show that the contributions of taxa in microbial associations are disproportionate to their abundances and that rarer taxa of microbial species play an integral part in shaping dynamics of microbial community interactions. Network-based meta-analysis revealed valuable insights into microbial community dynamics between healthy and disease phenotypes. We anticipate that the healthy and diseased microbiome association networks we inferred will become an important resource for human-related microbiome research.

The gut microbiome serves to provide a wide range of symbiotic functions, including metabolism, immune system development, and pathogen resistance[1]. While the gut microbiome plays an important role as a modulator of host health and disease, commensal colonizers are often susceptible to disruption, which has been shown to be associated with the development of disease states[2–4]. The advancement of sequencing technologies has fueled the rapid expansion of metagenomic data availability, enabling association studies between the human microbiome and various disease states[5,6]. While many microbiome studies rely on differential analysis to identify individual bacteria of interest between cohorts, the ability of network analysis to provide high level insights into global and local structures makes it an attractive approach to study the dynamic nature of microbial communities.

Metagenomic co-occurrence has been widely applied in metagenomic studies to construct microbiome networks and better understand microbiome community structures[7–10]. Features of metagenomic data pose several challenges to microbial co-occurrence analysis. Firstly, as sequencing technologies are not able to capture the true absolute microbiome abundance of samples, sequence abundances need to be represented as a proportion, rendering species abundance compositional by nature[11]. However, relative abundances, a common measure used to represent microbial abundances, is often considered a flawed metric to use in co-occurrence-based approaches due to the constant sum constraint, where assumptions of correlation metrics such as the independence between features are violated[12,13]. As relative abundances of species are dependent on the relative abundances of every other species present, abundance values of a given sample are no longer independent of each other when normalized to relative abundances. As such, alternative methods of normalization or transformation of raw abundance values remain necessary to compare species co-abundances across samples of varying sequencing

Luddy School of Informatics, Computing and Engineering, Indiana University, 700 N. Woodlawn Avenue, Bloomington, IN 47408, USA. ✉email: yye@indiana.edu

depths. Additionally, the use of compositionally aware association measures and methods are needed to handle the compositionality of microbiome datasets[14–16]. Various methods have been proposed to address the challenges of analyzing compositional data, and these methods that have been reviewed in detail[11,15,17–20]. Secondly, microbiome data is often subjected to issues of sparsity, where microbiome abundance matrices are zero-inflated due to heterogeneity within and between samples. Rare taxonomic species and/or insufficient sequencing depths contribute to the sparsity often seen in microbiome datasets[21,22]. The sparsity found in metagenomic datasets introduces challenges to log-ratio based transformation techniques used to handle compositionality. Additionally, correlations of sparse datasets can lead to strong spurious correlations[16,21]. Non-parametric and ranked-based correlation measures such as Spearman's Rho and Kendall's Tau are also susceptible to multi-way ties due to matrix sparsity and heavy-tailed distributions, and quickly deteriorate in presence of many zeros[13,23,24]. Thirdly, indirect correlations can often add noise to correlation-based interaction inference methods, where these indirect associations (e.g. spurious associations) can be driven by indirect species associations, batch effect, or environmental factors[10,16,25–27].

Despite the challenges of utilizing co-occurrence metrics on metagenomic datasets, a wide range of methods have been adopted, developed, and utilized to better understand microbial associations. In general, methods used to study microbial associations can be grouped into two categories: (1) traditional/classical correlation methods (e.g. Pearson, Spearman, Kendall's Tau), and (2) compositionally-aware methods. While compositionally-aware methods vary in their algorithms, they all seek to mitigate the confounding factors imposed by the current limitations of compositionality found in microbiome datasets. Compositionally-aware methods can be further sub-categorized into correlation-based methods (e.g. SparCC[28], CoNet[29], CCLasso[30]) and conditional dependence methods (e.g. SPIEC-EASI[25], Flashweave[26]) which try to differentiate between direct and indirect conditional dependencies. While the review and benchmarking of available methods' performance remains beyond the scope of this paper, the discussion surrounding the complexities of various microbial inference techniques have been reviewed at length[7,12,13,17,21,31–33]. As studies have previously shown, the results of networks generated from microbial association inference are largely dependent on the method used to infer the microbial interactions[12,13,31,33]. Methods of interaction inference vary largely between studies in terms of accuracy and precision, and no one existing tool is able to address all issues of biases or confounding factors[12,21,26,31,33].

Recently, various pipelines and tools have been developed to provide microbiome network-based analysis, including NetCoMi[32] and iNAP[34]. As there remains a lack of a community consensus and gold standard to evaluate the performance of methods used to infer microbial co-occurrence networks, users are largely left to decide the method of inference and it remains imperative for users to understand statistical considerations, such as those mentioned above, when deciding downstream methodology. Here in this study we utilize SPIEC-EASI[25] as the association method for microbial association inference, considering that this method takes into account the compositionality of microbiome data to mitigate potential indirect associations. In conjunction with utilizing a compositionally aware correlation method, we employ various pre-processing steps to help mitigate challenges commonly associated with metagenomic correlation-based analyses.

Variation between datasets come not only from intra-sample heterogeneity, but also different preprocessing and post-processing methods used between studies. The lack of consensus in computational methods, including annotation, quantification, preprocessing, and association methods makes comparison of findings between studies difficult. Despite the significant progress in methods development for compositionality-robust association methods and known issues with traditional correlation-based methods, traditional correlation methods (e.g. Spearman) still remains the most widely used type of association metric. The slow adaptation of compositionally aware methods for metagenomic data remains multi-factor and can likely be attributed to the exponential increase in computational requirements of compositionally aware methods, as well as legacy effect where researchers adopt the methods used in previous studies.

Here in this study, we utilize a large collection of healthy and disease gut metagenome datasets to preform a meta-analysis using microbiome association networks by re-analyzing and standardizing the analysis approach. We note that the datasets used in this study were originally compiled in[35], where Gupta et al. used these datasets to identify 7 health-prevalent and 43 health-scarce bacterial species, from which they developed a Gut Microbiome Health Index (GMHI) for evaluating health status based on the species-level taxonomic profile of a stool microbiome sample. While the meta-analysis preformed by Gupta et al. was able to demonstrate improved patient stratification between healthy and diseased microbiomes compared to common alpha-diversity measures, remaining misclassification between samples demonstrates the complexities of defining a stratification criterion owing to our limited understanding of gut microbial ecology and their relation to human health. To build on these efforts, we constructed microbial association networks utilizing a subset of samples used by Gupta et al. Furthermore, we focus our efforts in analyzing diseases individually, in contrast to a disease-agnostic approach utilized by Gupta et al., to better characterize individual disease microbial community traits. By doing so, we expand the existing literature by uncovering microbiome community associations and community assembly dynamics within and between healthy and diseased microbial communities in an effort to identify features to help stratify disease states and potential microbial risk factors beyond individual species. Additionally, to better understand community interactions across phenotypes, we also introduce a new measure termed 'module resilience' to study microbial community modules retention across microbial interaction networks.

Materials and methods

Datasets and preprocessing. A curated list of sample accession numbers from publicly available human gut metagenome datasets was gathered from Gupta et al.[35] to be used this study. Gupta et al. used a total of 4347 samples from 78 different study accessions, with samples spanning 13 different phenotypes. In our analysis, we only included 4143 samples from 10 phenotypes: Healthy, advanced (colorectal) adenomas (AA), atheroscle-

Phenotype	# of studies	# of samples
Healthy	29	2568
Advanced (colorectal) adenoma (AA)	2	82
Atherosclerotic cardiovascular disease (ACVD)	1	152
Colorectal cancer (CRC)	4	254
Crohn's disease (CD)	4	107
Obesity (OB)	15	324
Overweight (OW)	16	232
Rheumatoid arthritis (RA)	1	92
Type 2 diabetes (T2D)	3	236
Ulcerative colitis (UC)	3	96
Total	78	4143

Table 1. Summary of gut microbiome datasets used in downstream gut microbiome association network analysis.

rotic cardiovascular disease (ACVD), colorectal cancer (CRC), Crohn's disease (CD), obesity (OB), overweight (OW), rheumatoid arthritis (RA), Type-2 diabetes (T2D), ulcerative colitis (UC). Samples from the following phenotypes included in Gupta et al., impaired glucose tolerance (IGT), symptomatic atherosclerosis (SA), and underweight (UW) were excluded from downstream analysis due to low sample count. Samples were downloaded from NCBI Sequence Read Archive via SRA Toolkit's fasterq-dump.

A summary of samples used in this study can be found in Table 1. The works of Gupta et al. focused the meta-analysis of gut microbiome species to develop a health status index that utilizes species-level gut microbiome profiling to stratify between microbiome health states. While we utilized a similar dataset to Gupta et al.'s study, there are several notable differences between our analysis approach. Firstly, we selected a Kraken2+Bracken approach for microbial quantification and taxonomic assignment due to its superior performance compared to marker gene based methods as highlighted in a recent benchmark of metagenomic classification tools[36], where marker gene based methods ranked among the lowest among assessed tools in terms of precision and recall for species classification and lowest proportion of abundance quantified at species-rank. Secondly, while maintaining similar study accessions, we insured that all run accessions downloaded focused on available paired-end reads with the largest available spots rather than utilizing a mix of single-ended and paired-ended reads. Finally, our meta-analysis focuses on species co-occurrences and network-based approaches rather than focusing on the prevalence of species-level abundances between samples, and the bacterial networks resulted from our analyses can be used by other researchers for different research purposes.

Samples were processed to remove low quality reads and Illumina adapters using Trimmomatic (v0.39)[37] with parameters SLIDINGWINDOW:4:20 LEADING:20 TRAILING:20 MINLEN:60. Trimmed samples were then mapped to the human genome assembly GRCh38 (hg38) using bowtie2 (v2.4.4)[38] to remove possible human read contamination from the metagenome samples. All remaining unmapped metagenomic reads were kept for downstream analysis. Additionally, low read count samples that were less than 1M reads were discarded from this analysis to prevent inclusion of under-sampled genomes. Distribution of the filtered reads can be found on Supplementary Fig. 1. Following filter and trimming of samples, a total of 4143 out of the original 4347 samples were retained for downstream analysis. A complete list of accessions used in this analysis can be found in the GutNet repository.

Microbiome taxonomic assignment and abundance quantification. Taxonomic assignment and species abundance quantification were preformed using Kraken2 (v2.0.8)[39]. The pre-built 'Standard' Kraken2 database (version k2_standard_20201202) maintained by the authors of Kraken2, built on December 2, 2020, was used as taxonomic references (https://genome-idx.s3.amazonaws.com/kraken/k2_standard_20201202.tar.gz). The 'Standard' Kraken2 database was built using RefSeq reference genomes, including references from archaea, bacteria, viral, plasmid, human, and UniVec_Core databases. Only archaea and bacterial counts were retained for downstream analysis. Kraken2 prokaryotic taxonomic assignments and abundances were then re-estimated with Bracken (v2.6.2)[40] for species-level re-estimation of abundances. Samples were aggregated into their representative disease phenotype to construct species level read abundance matrices.

Species abundance processing and filtering of sparse taxa. One of the challenges in dealing with metagenomic data for co-occurrence inference is the sparsity of metagenomic data. This sparsity can be attributed to a multitude of factors (e.g. sequencing depth, sample heterogeneity) and can cause spurious correlations and false-positives in statistical methods[12,14]. To address some of the issues caused by matrix sparsity, we employed a method of filtering based on species prevalence similar to those suggested by[12,41]. To determine the level of species prevalence to filter, we empirically evaluated the species:species-prevalence distribution within our datasets to determine a species prevalence threshold that minimized zero-inflation while retaining majority of species locally observed within each respective phenotypic group (Fig. 1). Evaluating this distribution, we determined that a 50% prevalence threshold was a conservative threshold and also consistent with the suggestions of Weiss et al.[12]. Within-phenotype species-level abundance matrices were then filtered to remove low

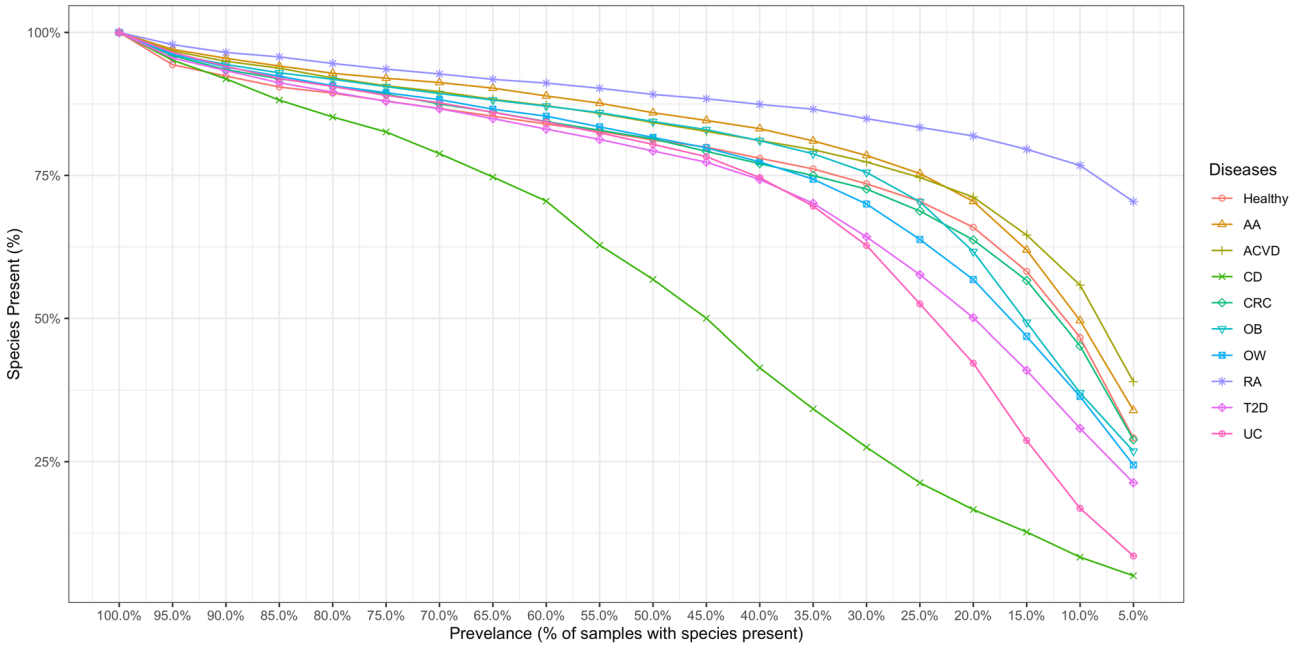

Figure 1. Identified species retained decreases with increasing prevalence threshold. The X-axis represents the prevalence threshold to filter species at in increments of 5%. The Y-axis represents the proportion of species represented in a given disease phenotype abundance matrix.

prevalence taxa below a 50% sample prevalence threshold, and filtered abundance matrices were then used for all downstream correlation based analyses.

Microbiome association. Using prevalence filtered Bracken reads count abundance matrices, species-level associations were inferred for each disease abundance matrices respectively. SPIEC-EASI[25] was selected as the association method for microbial association inference due to the method accounting for compositionality of microbiome data and potential indirect associations. SPIEC-EASI was run using the 'MB' method, a neighborhood selection method developed by Meinshausen and Bühlmann[42] used to infer sparse inverse covariance matrices from a network. SPIEC-EASI has been found to preform well in comparison with other association methods, and thus selected to be used in this analysis[14,16,26].

Microbiome network construction. Microbiome co-occurrence network were constructed from association values computed using SPIEC-EASI, where values were filtered with a 0.1 absolute association value threshold. Network vertices were defined as prokaryotic species for species-level networks; vertices and node are used synonymously throughout. An undirected edge was constructed between two vertices if a significant association between two given vertices was inferred. Edge weights range between [− 1, 1], where positive edges represent a positive association and negative edges represent negative associations. It should be noted that edge weights of conditional dependence methods cannot be directly compared to correlation based metrics and are not directly proportional even though their values are assessed on the same scale (e.g. Pearson, Spearman, SparCC)[25,26]. Networks were visualized through Gephi[43] using Force Atlas 2 layout. All singleton nodes without edges were removed from the network.

In addition, a consensus network was constructed to analyze Proteobacteria interactions among the disease networks. Given an edge, if any vertices within that edge had an annotated Genus as Proteobacteria, the edge was kept. Utilizing all remaining edges, a consensus network for each disease was built, where the edge weight was equivalent to the number of networks containing each respective edge.

Community module detection. Many methods developed for community module detection in network systems are based off of undirected, unsigned, and positive networks. However, methods for signed module detection remain largely under-explored. In many cases, negative edges are simply discarded or ignored. However, as microbiome interactions are highly-dynamic and involve not only positive interactions, it is important to maintain the use of signed interactions when possible. To address this challenge, we utilized the Leiden algorithm[44], which attempts to extend on the works of the Louvain algorithm[45]. The Louvain algorithm can sometimes have badly connected communities, whereas the Leiden algorithm guarantees that communities are well connected and locally optimized. The Leiden algorithm consists of three steps, first it performs a local moving of nodes, second it refines partitions, and lastly the aggregation of the network based on the refined partitions. The Leiden algorithm takes advantage of local moving procedure and is able to split clusters rather than only merging them as in the Leiden algorithm. Additionally, the Leiden algorithm is able to handle negative edge weights.

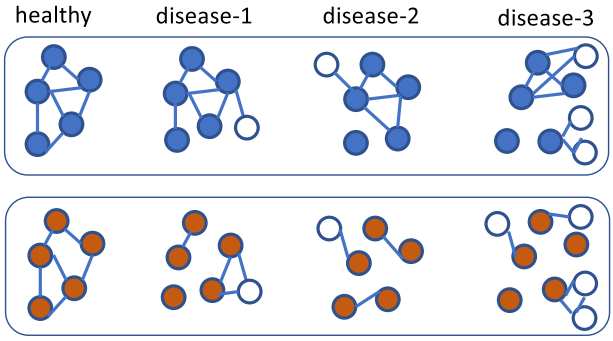

Figure 2. A toy example demonstrating module resilience. The healthy network contains two modules, the blue module containing five species shown as blue nodes and the red module also contains five species shown as red nodes. The blue module, whose composite nodes tend to remain in the same module across diseased networks, has higher resilience (resilience score = 0.8) than the red module (resilience score = 0.4).

Module resilience. We proposed a resilience score to approximate the tendency of modules of gut bacterial species detected from the healthy microbiome network to remain in the same community in the gut microbiome associated with different diseases. Given a module i found in healthy network containing r_i species, for each diseased network our approach finds the module in the diseased network j that contains the most members of the r_i species (denote as d_j) (so $\frac{d_j}{r_i}$ indicates the tendency of the species in module i staying in the same community in diseased network j). The resilience of module i is defined as the median of $\{\frac{d_1}{r_i}, \frac{d_2}{r_i}, \ldots, \frac{d_K}{r_i}\}$, where K is the number of diseased networks ($K = 9$ in this paper). For example, module i contains 20 species, and 16 out of these 20 species are found in a module in the microbiome network for disease j (the remaining 4 species are found elsewhere), then $d_j = 16$ and $\frac{d_j}{r_i} = 0.80$. Assume $\frac{d_j}{r_i}$ is 0.80, 0.90, 0.60, 0.70, 0.85, 0.75, 0.90, 0.35, 0.40 for $j = 1, \ldots, K$, respectively, module i has a resilience score of 0.75 (the median). See Fig. 2 for an illustration of module resilience. While this analysis was able to identify modules that were likely to be resilient to change, it does not provide information in regards how necessary the module was in regards to microbiome health nor does it identify 'core' microbiota, instead it shows how likely microbial species were to consistently form community modules across networks.

Availability of the programs and inferred networks. All network (GML) files, bioinformatics workflows, and analysis scripts produced as part of this study can be found in a GitHub repository https://github.com/mgtools/GutNet. Sample run accession numbers and associated study accession for all publicly available stool metagenome samples used in this study are available in the repository.

Results

Microbiome composition and sparsity problem. The total number of species annotated in all datasets was 6463, spanning 4143 samples (see Table 1). When agglomerated at the Phylum level, we unsurprisingly found that Bacteroidetes, Firmicutes, Proteobacteria, Actinobacteria, and Verrucomicrobia were the 5 most dominant Phylum, with Bacteroidetes and Firmicutes dominating over 80% of the total relative abundance (Fig. 3). This distribution of observed top Phyla is in line with previous studies that found similar distributions of top Phylum-level abundances in human gut microbiome[46–48].

It has been shown that sparsity of microbial datasets affect correlation methods, and often result in spurious correlations. To address this issue, various explorations have proposed the use of filtering rare microbial taxa[12,41]. Filtering species with low prevalence reduces the zero richness within datasets and helps resolve some of the statistical artifacts imposed by sparse datasets. Before deciding on a prevalence threshold, we evaluated the effect of imposing a prevalence threshold on microbial taxonomic distributions (Fig. 1). In most disease abundance matrices, the observed species present gradually decreases until approximately 65% prevalence, where thereafter the number of species post-filtering sharply decreases; the CD abundance matrix was the exception, where CD had a more linear relationship in terms of percent of species retained and percent of prevalence filtered.

We show that for all abundance matrices (except CD), a prevalence filter of 50% as suggested by[12] will result in a reduction in the number of species between 10.85–20.73% relative to the unthresholded datasets; with the exception of the CD dataset which will incur a 43.18% reduction in number of species observed after thresholding at 50% prevalence. Given the marginal differences in the number of species removed at prevalence values less than 50%, except for CD, we decided that a 50% species prevalence threshold was acceptable. Additionally, for the CD abundance matrix we decided that the trade-off of reducing sparsity was enough to warrant the loss of species present within the dataset, thus followed a 50% threshold for prevalence on all abundance datasets.

Assessment and comparison of microbiome ecological diversity in phenotype specific microbiomes. To evaluate the alpha-diversity between healthy and diseased microbiome datasets, we utilized the Shannon diversity index and species richness (observed number of different species) measures per each phenotype (Fig. 4). For the alpha-diversity based on the species richness, we found that healthy datasets had a statistically significant different distribution compared to diseased datasets in terms of species richness (Fig. 4A; two-

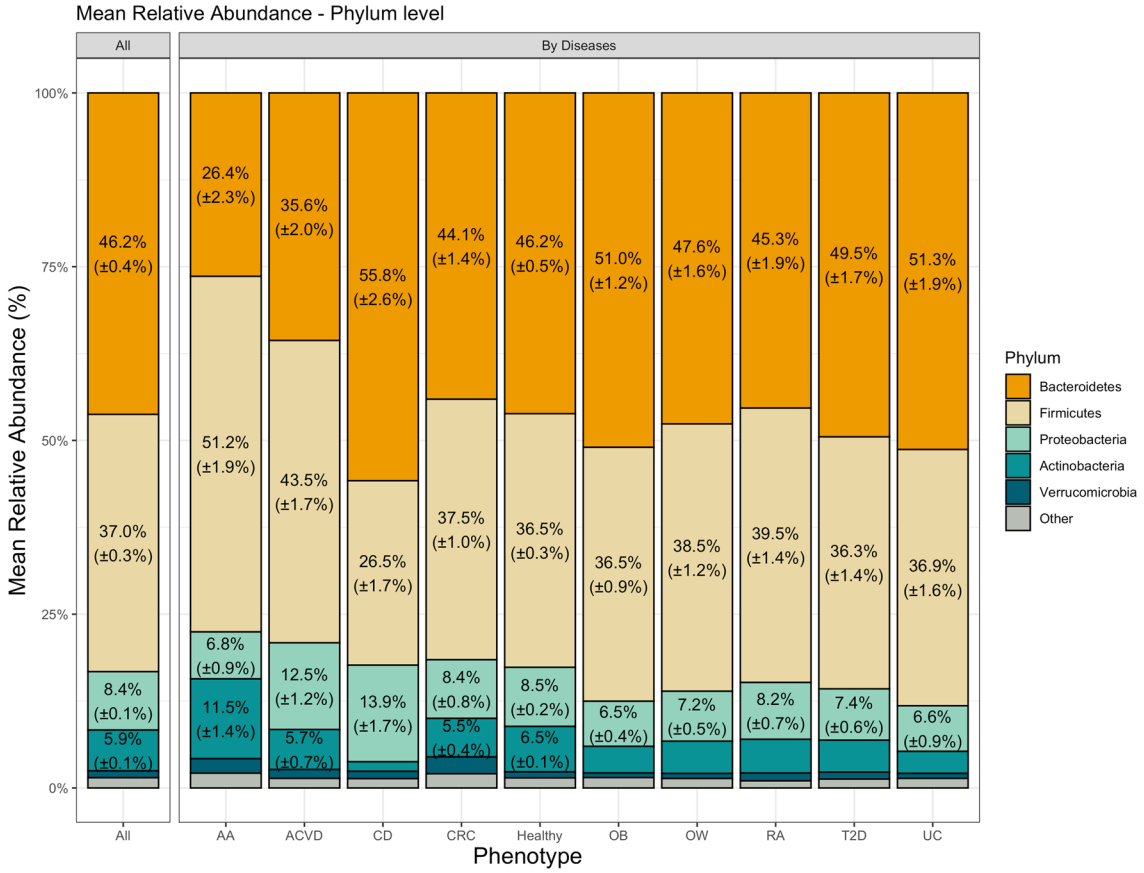

Figure 3. Mean distribution of species found within metagenome datasets by phenotype, agglomerated at the Phylum level. Numbers within each bar represent the mean relative abundance, accompanied by its standard error; only values above 5% are shown.

sided Mann–Whitney U test, p-value $= 2.9e{-}4$). Additionally, when testing the statistical difference between the healthy datasets versus each disease dataset individually, 8 out of 9 diseased phenotypes (i.e. AA, CD, CRC, OB, OW, RA, T2D, and UC) were found be statistically significant (Fig. 4B; two-sided Mann–Whitney U test, p-value < 0.05). Shannon diversity measures between healthy and diseased datasets also showed a statistically significant different distribution (Fig. 4C; two-sided Mann–Whitney U test, p-value $= 1.6e{-}12$). Comparison between healthy and individual disease phenotypes also showed 5 out of 9 disease phenotypes (i.e. ACVD, AA, CD, T2D, UC) to be statistically significant (Fig. 4D; two-sided Mann–Whitney U test, p-value < 0.05). Observations of significant differences in alpha diversity measures between healthy and diseased datasets are in line with previous studies that have used alpha-diversity measures as an indicator of disease-associated microbiome dysbiosis[49,50].

For beta-diversity analysis, we used ordination plots to summarize the microbiome community data of healthy population and individuals with diseases. We used Bray-Curtis dissimilarity as the distance measure between the datasets, and used both t-SNE and NMDS approaches for dimensionality reduction. In the 2-dimensional ordination space shown in Supplementary Fig. 2, samples with similar microbial compositions are close in the plots. The ordination plots show that samples did not cluster at the phenotypic-level, indicating that there is no discernible structure to microbiome abundance profiles that stratifies diseases purely based on taxonomic features. For comparison, the PCoA plot of the samples from Gupta et al. (Figure 3d in[35]) also showed no clear clusters of the samples according to phenotypes, but in their study, an ANOSIM test showed weak difference between among- and within-group dissimilarities.

Microbial association network and resilient modules. To better understand microbiome associations and microbial community interactions in healthy and diseased gut microbiomes, we identified microbiome community modules within each microbiome network (Supplementary Figs. 3–12). Co-occurrence networks were constructed for each phenotype, and community modules in each network were identified utilizing the Leiden algorithm. We compared the modules identified in the different microbiome networks to study the community module stability. By understanding the module resilience, we were able to identify microbiome community modules that were resilient to change, and identify species of bacteria that were more likely to be associated to each other regardless of the environment.

In our analysis, we were able to identify several modules of high module resilience (Fig. 5). In many cases, modules of high resilience were populated by members of the microbiota within the same clade. These include

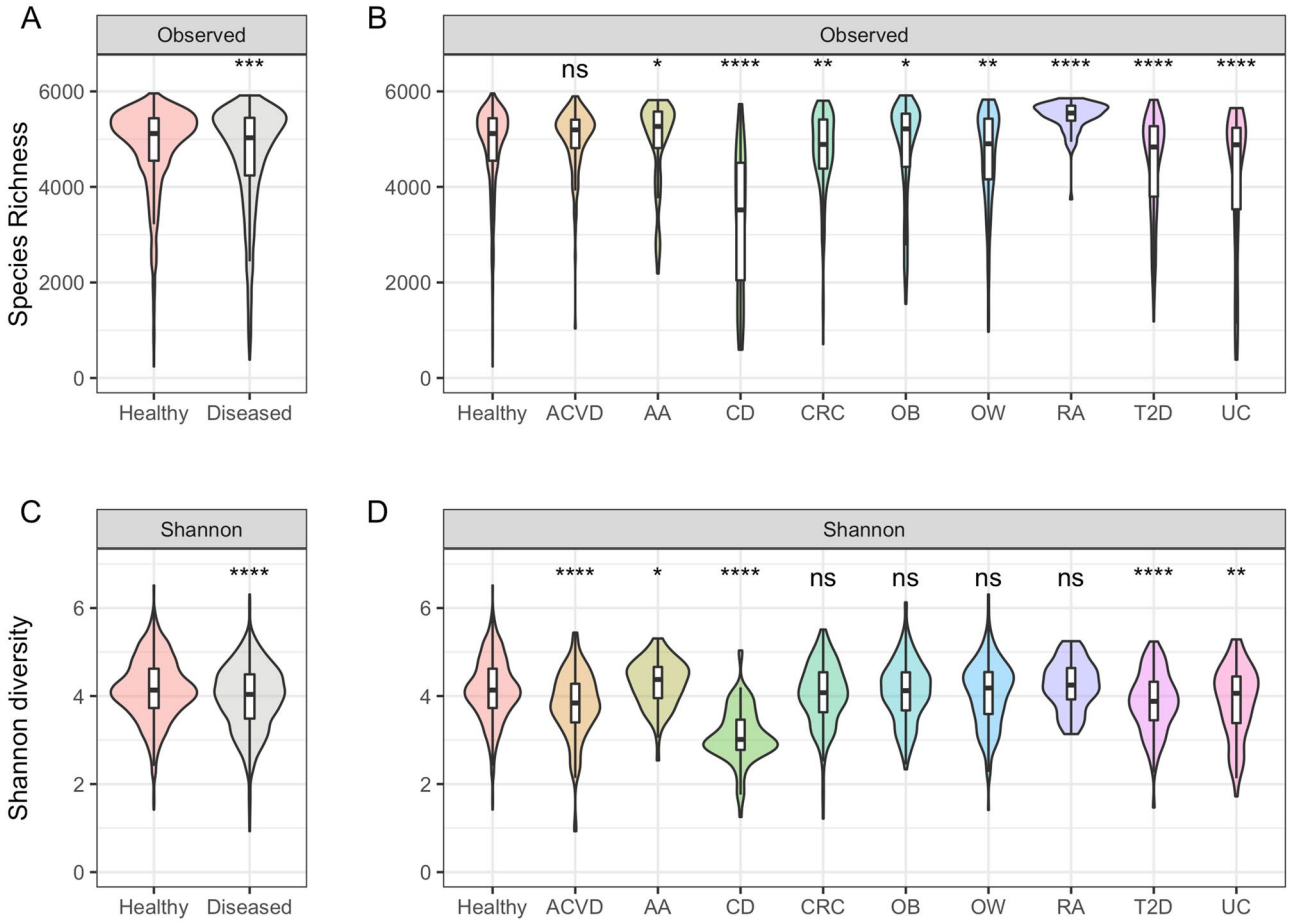

Figure 4. Alpha-diversity comparisons between datasets. (**A**) Species richness plot between healthy and diseased datasets, (**B**) species richness plot comparison between all phenotypes, (**C**) Shannon-diversity plot between healthy and diseased datasets, and (**D**) Shannon-diversity plot between all phenotypes. Two-sided Mann–Whitney *U* test was used to compare respective disease datasets against the healthy dataset. The p-value significance are shown above violin plots: ns (non-significant; p-value > 0.05), *(p-value < 0.05), **(p-value < 0.01), ***(p-value $< 1e{-}3$), ****(p-value $< 1e{-}4$).

modules which were found to be *Streptococcus*-rich and *Escherichia*-rich at the Genus level, as well as Actinobacteria-rich and Proteobacteria-rich at the Phylum levels. We note that the *Streptococcus*-rich module contains *S. anginosus*, *S. australis*, *S. gordonii*, *S. sanguinis* and *S. vestibularis* that were considered to be health-scarce species previously by Gupta et al.[35]. Additionally, we also found modules with a mixture of Phyla that also exhibited high resilience, suggesting that resilience of modules may include both taxonomically assortative communities and those of mixed communities. While module resilience does not provide context as to why certain modules of microbial associations were retained through both healthy and diseased networks, it can help us better understand the underlying community structure and generate candidates for downstream hypothesis testing (Fig. 5).

Contributions of taxa in microbial association networks are disproportional to their abundances. By examining the species (i.e., the nodes) and their interactions (i.e., the edges) in the microbial association network, we can study their contribution to microbial community assembly. Analyzing the nodes of constructed association networks, we found that the top Phyla in each association networks comprised of Proteobacteria, Firmicutes, Actinobacteria, Bacteroidetes, Euryarchaeota, and Cyanobacteria (Supplementary Fig. 13). The most abundant interacting Phylum was that of Proteobacteria, which represents 42.37% of the total nodes found in the SPIEC-EASI association networks. This is in contrast to Bacteroidetes and Firmicutes which together only represented 26.89% of the total nodes found in SPIEC-EASI association networks although they together represented $> 80\%$ of the mean total relative abundances (Fig. 3). The discrepancy of the prevalence of the species and their contribution to the association networks suggests the importance of studying bacterial interactions and networks. Our findings here are in contrast to those found in Gupta et al.[35] where Firmicutes comprised a significant portion of species found in their analysis to be enriched in disease samples; Firmicutes comprised of 37 out of 50 species (74%) used to compute the GMHI score. This contrast suggests that beyond differential microbial abundances, microbial interactions can also play a pivotal role in stratifying microbiome disease states.

Taking a closer look at microbial interactions of gut microbiomes between healthy and disease datasets, we analyzed the Phyla distribution of edge associations within each network. Similarly to network nodes, Phyla

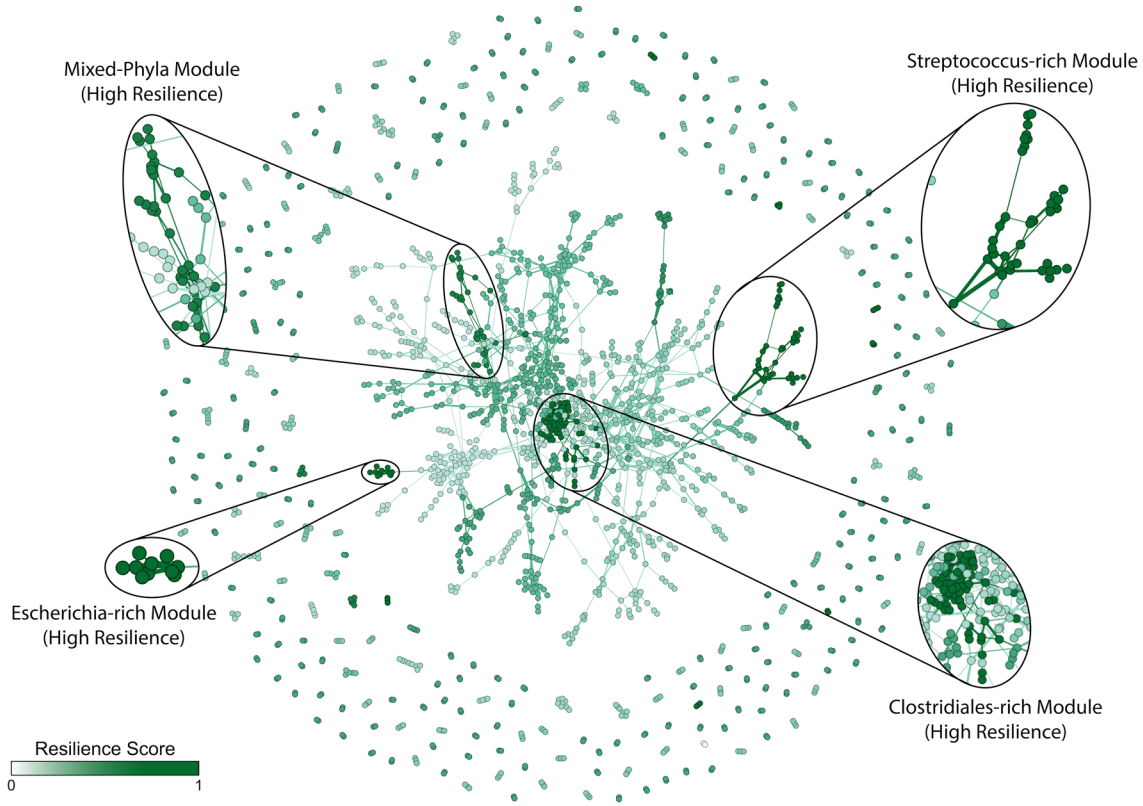

Figure 5. Microbiome association network, colored by module resilience. Module resilience scaled between [0,1] with lighter color modules represent lower module resilience, and darker color modules represent higher module resilience.

distribution of edges also did not show preference to Bacteroidetes and Firmicutes despite the dominant proportion of both Phyla in terms of relative abundances. As many of the interaction edges between microbial members lie between less populous Phyla, this highlights the importance of rarer species of the microbiome.

Differentiating between positive and negative edges in the network, we analyze the differences within and between the microbiome networks (Fig. 6). Of notable observations, both Bacteroidetes–Bacteroidetes and Firmicutes–Firmicutes interactions were enriched in healthy populations compared to their diseased counterparts. While Firmicutes and Bacteroidetes did not exhibit drastic mean abundance differences in most disease datasets compared to the healthy dataset, the decrease in self-Phylum interactions may suggest that Firmicutes–Firmicutes and Bacteroidetes–Bacteroidetes interactions play an important role in maintaining gut homeostasis. Additionally, we found that Proteobacteria–Actinobacteria associations were enriched in disease networks compared to the healthy network and may be a signature of microbiome dysbiosis.

Proteobacteria interactions enriched in disease association networks. Previous studies have found that microbial abundances of Proteobacteria species were enriched in diseased microbiota and also have also proposed that Proteobacteria may be a signature of disease[51,52]. While our results do not show consistent increase in mean relative abundance of Proteobacteria across all diseased datasets (only ACVD and CD datasets had a mean relative abundance greater than the healthy dataset; Fig. 3), we observed that Proteobacteria participation in interactions (i.e., network edges) were significantly enriched in all disease networks. Proteobacteria was found to be the most dominant phylum in terms of network edge participation, where Proteobacteria was part of either one or both vertices in a given network edge. On average, Proteobacteria participated in about 59% of the interactions in the microbiome association networks (healthy and diseased). Interestingly, the healthy network was identified as an outlier among networks (following Tukey's method of outlier detection) with 33.88% of the interactions involving Proteobacteria (Fig. 7A).

Taking a closer look at Proteobacteria edges within our networks, we found that non-disease Proteobacteria interactions were often connecting modules pre-dominantly interconnected with Proteobacteria containing edges that were also found in the healthy network (Fig. 7B). This result shows that beyond microbial co-occurrences commonly shared between healthy and diseased networks, the diseased networks also contain disease-only edges that greatly interconnect Proteobacteria species compared to the healthy network. Additionally, majority of Proteobacteria containing edges within the consensus network were filtered out, being observed in less than 5 networks, suggesting that many of these Proteobacteria connections are not universal across all diseases. Together, this may suggest that the enrichment of Proteobacteria edges observed in disease networks are contributed by

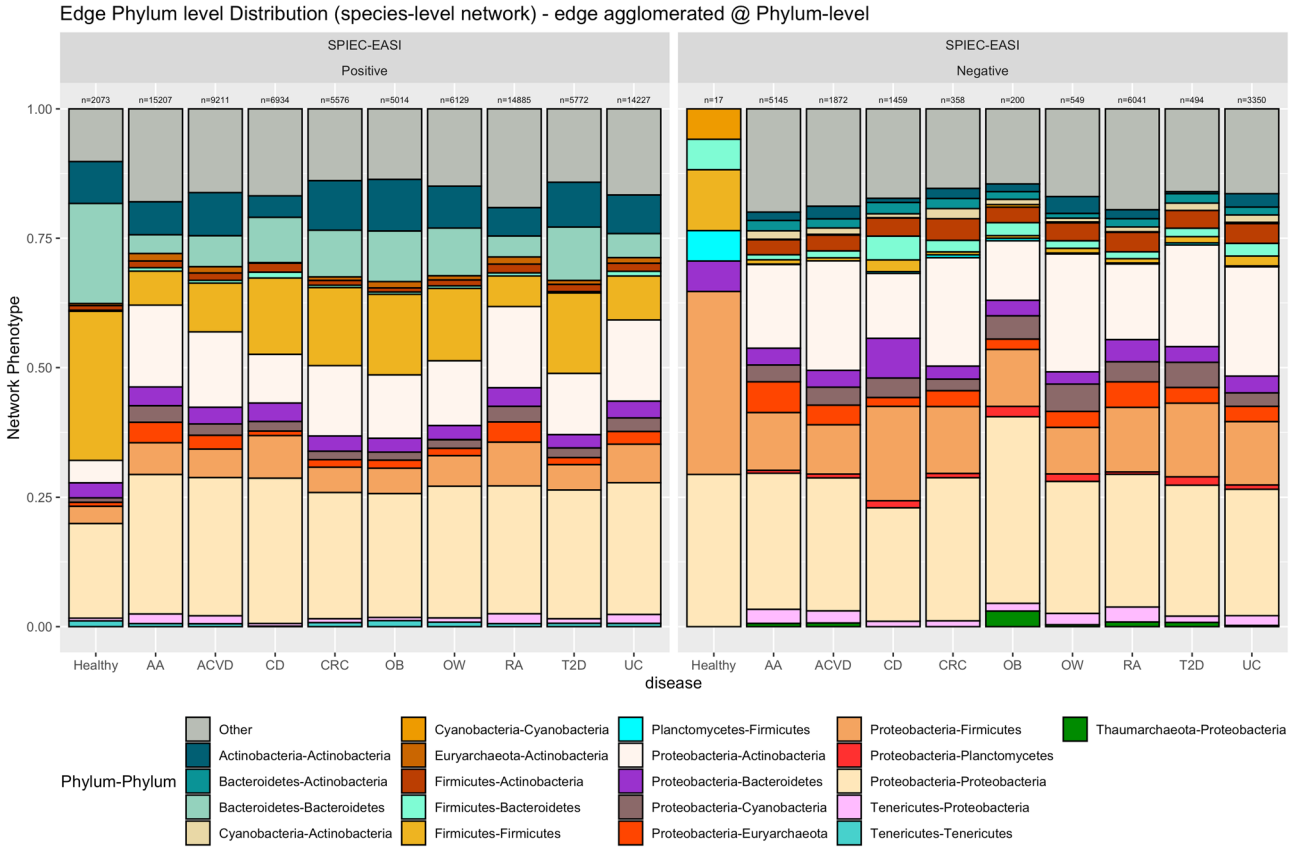

Figure 6. Taxonomic distribution (at the Phylum level) of the species involved in microbiome association networks by phenotype. (left) Positive edge distribution stacked barplots, (right) Negative edge distribution stacked barplots.

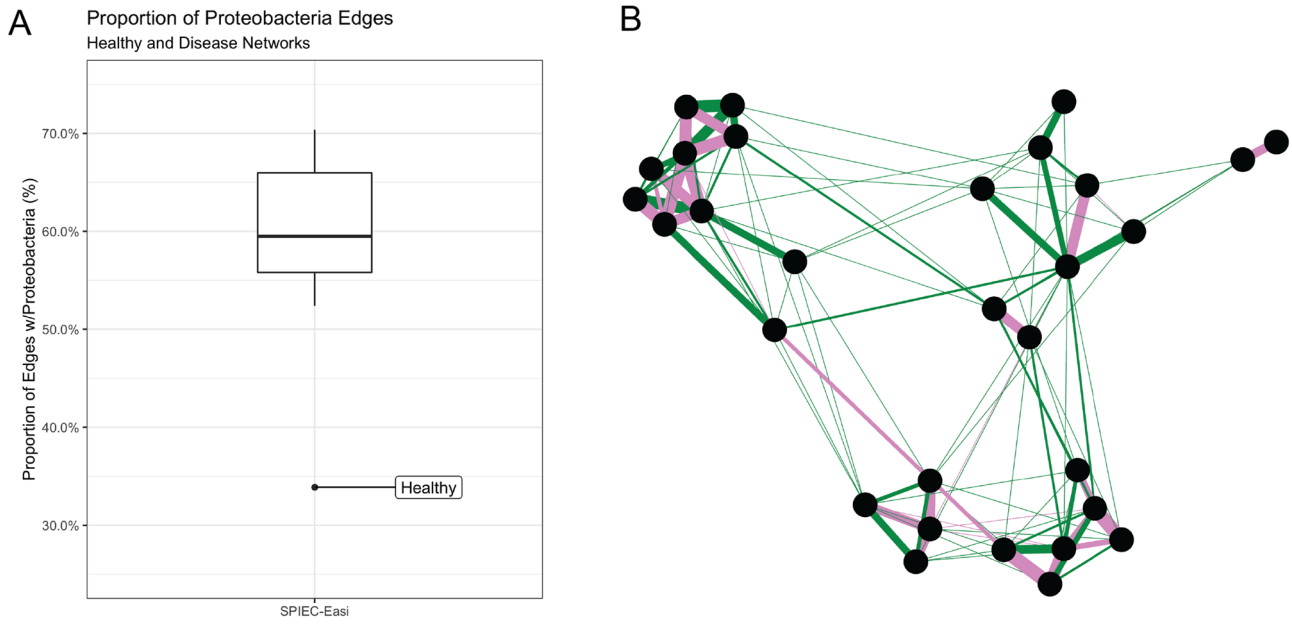

Figure 7. Proteobacteria interactions in microbial networks. (**A**) Distribution of the fractions of interactions that involve Proteobacteria in healthy and diseased microbiome association networks. In the boxplot, the Y-axis represents the proportion of interactions involving Proteobacteria. (**B**) Subgraph containing the largest Proteobacteria module found in consensus network. Consensus network contains edges shared between 5 or more disease networks. Green edges represent edges that are not found in the healthy network, while pink edges represent edges in the consensus network that are also found in the healthy network. Edge weight is scaled by the count of networks that a given edge is observed in.

rare disease-specific edges, and provide greater interconnectivity between Proteobacteria containing edges that would be otherwise be considered loosely connected when compared to the healthy network.

Discussion

Here in this meta-analysis of gut microbiome datasets, we report patterns of microbiome interaction within and between healthy and diseased microbiomes through the use of microbiome association networks. Our analysis showed that rare taxa of microbiome datasets can play a disproportionate role in microbiome interactions relative to their taxonomic abundances. While Bacteroidetes and Firmicutes were found to comprise a majority of the microbiome abundances in all microbiome phenotypes, the proportion in which Bacteroidetes and Firmicutes participated in significant network associations in terms of both nodes and edges were unproportional to their high relative abundances. Instead, majority of the significant edges within the microbiome association networks were composed of rarer taxa. This contrast supports previous studies that suggests that rare species may play an over-proportional role in microbiome community dynamics compared to their more abundant counterparts[53–55].

In our observations, we also found several notable differences between healthy and diseased microbiome networks. These observations include an enrichment of Bacteroidetes–Bacteroidetes and Firmicutes–Firmicutes interactions within the Healthy Network and enriched Proteobacteria–Actinobacteria interactions in Diseased Networks. While it is unclear if these differences in association are causal or a result of a diseased state, these differences in interactions highlight dysbiosis in diseased microbiome association networks and can be used as potential markers. Additionally, Diseased network edges were found to be highly enriched for Proteobacteria compared to the Healthy network. The Healthy network had a significantly lower proportion of Proteobacteria participation in association networks compared to Diseased networks, and suggests that increased Proteobacteria interactions with other members of the microbiome may be a hallmark of microbiome dysbiosis. Many of the features identified in this study that stratify between healthy and disease networks were found to be consistent across disease networks, suggesting that these features are not disease-specific but general markers of dysbiosis and features of diseased gut microbiota.

Additionally, by identifying community modules within both Healthy and Diseased networks, we were able to identify community modules that were resilient to change and the community interactions that were likely to be retained across different microbiome association networks regardless of phenotypic association. While these modules do not necessarily represent a 'core' microbiome associated with a particular phenotype, these resilient modules help us better understand the underlying microbiome community structure that is shared between phenotypes. Investigation into better understanding of microbiome community structure and assembly dynamics can help future efforts in modulating the human gut microbiome. Module resilience highlights the advantages of meta-analysis, and utilizing standardized approaches so that cross-disease and cross-study analysis can be generalized across datasets to help us better understand microbiome dynamics spanning across diseased states.

While this analysis did not include all possible studies or diseases, this study highlights the benefits of re-analyzing studies with standardized procedures so that results can be generalizable and compared between datasets. That being said, there still remains much limitations to microbiome meta-analyses and microbiome interaction as a whole. In particular, as there often lacks widely accepted reference standard and adopted protocol, methods and techniques utilized to analyze microbiome data is widely left open to interpretation and researchers can only inform themselves of the nuances between methods and select the method that best fits their data, needs, and available resources. Issues of possible variation and confounding factors such as experimental or sequencing artifacts, environmental factors, batch effect, differences in taxonomic annotation and quantification methods, technical artifacts highly limit robust downstream analysis. To mitigate some of the potential issues of confounding factors (e.g. species-level annotation error, batch effect between studies, variation between study design and patient selection), we focused majority of our analysis by agglomerating at the Phylum level, but acknowledge there remains much more to be explored at lower taxonomic levels. There remains an increasing need for gold standards to be developed so that tools and methods can be benchmarked and evaluated to establish standardized protocols. Future efforts in development of experimental and computational methods are necessary to address issues of microbiome compositionality.

For network visualization, we utilized Force Atlas 2 which we note that the network layout depends on initial state of coordinates and can become stuck at local minimums[56]. While Force Atlas 2 may have certain limitations that need to be mindful of when interpreting visual representations of generated networks, all results here reported were derived from computational network models, and thus we believe that the findings of this work will not be impacted by network layout limitations if they are present.

We utilized the method of filtering for species prevalence as a means to mitigate potential statistical challenges resulting from sparse metagenomic abundances. While the recommendations from Weiss et al.[12] and Cao et al.[41] have suggested such prevalence filtering as an effective means to mitigate these challenges, both initially were based on 16S sequencing. However, their recommendations were made to address issues of sparsity and its influence on analysis of microbiome datasets, and thus their recommendations extended beyond 16S sequencing to include sparse abundance matrices also commonly found in metagenomic datasets. In fact, the practice of filtering for species prevalence is also commonly used to filter metagenomic sequencing results as a means to account for the same statistical issues, and remains particularly critical in correlation based analyses. Examples include those of Milanese et al.[57] that suggests the use of prevalence filtering from metagenomics abundance matrices to mitigate potential spurious correlations between low-prevalence, and Llyod-Prince et al.[58] that utilizes prevalence filtering to reduce affects of zero-inflation in metagenomic abundance matrices. Filtering methods will inevitably filter out species that are true-positives and there remains a possibility that some of these filtered species may play an integral part in influencing a given microbiome state. However, by utilizing a prevalence filtering method rather than an abundance filter, species that are observed homogeneously in within-phenotype microbiomes are

retained, including low abundance species. Without filtering for species prevalence, correlation based analyses risk the inclusion of spurious correlations that do not reflect true correlation but rather statistical artifacts.

Despite these limitations, our results uncovered features of microbiome association between healthy and diseased cohorts that may help future efforts in understanding alterations of the gut microbiome that may be associated with diseased states. For example, among the health-prevalent and health-scarce species identified by Gupta et al.[35], three health-prevalent Bifidobacterium species (*B. adolescentis*, *B. angulatum*, and *B. catenulatum*) and one health-scarce Bifidobacterium species (*B. dentium*) were found in all 10 healthy and disease association networks we derived, and it would be interesting to examine the interactions between these Bifidobacterium and other species in the networks and the differences across networks. While it is not possible to assess and benchmark the wide availability of microbial association methods, standardizing the protocols and processing steps of data analysis will help future efforts to uncover features that warrant further investigation. Here we provide all microbial association networks produced as part of this study as a resource for future efforts in studying microbial associations. By preforming meta-analyses, results of individual studies can reach beyond itself and assist in contextualizing new results through expanding insights in comparison to other studies. Nevertheless, computational microbiome association methods remain insufficient by themselves to identify causal interactions. Association analysis can only serve as a starting point to reduce the search space and identify potential candidates for downstream hypothesis testing and experimental validation.

Conclusions

We proposed a pipeline for microbiome association network inference that incorporates the recent advances in network inference approaches that can deal with sparse compositional data and tease apart indirect vs direct interactions. Through meta-analysis of inferred networks, we were able to identify network-associated features that help stratify between healthy and disease states. By focusing our analysis on microbial networks, we show that microbial interactions can extend approaches to stratify between microbiome associated disease phenotypes beyond differential abundances. The findings of this study add to the body literature to inform future efforts in microbiome related disease stratification efforts as well as efforts to better understand microbial community interactions. We made available the inferred healthy and diseased microbiome association networks in a standard network format and we anticipate that they will become an important resource for human-related microbiome research.

Data availability

All code, metadata, and graph files generated as part of this study is available in the GutNet repository located at https://github.com/mgtools/GutNet.

Received: 22 February 2022; Accepted: 17 October 2022
Published online: 19 October 2022

References

1. Koskella, B., Hall, L. J. & Metcalf, C. J. E. The microbiome beyond the horizon of ecological and evolutionary theory. *Nat. Ecol. Evol.* **1**(11), 1606–15 (2017).
2. Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**(7418), 55–60 (2012).
3. Lane, E. R., Zisman, T. L. & Suskind, D. L. The microbiota in inflammatory bowel disease: Current and therapeutic insights. *J. Inflamm. Res.* **10**, 63 (2017).
4. Shreiner, A. B., Kao, J. Y. & Young, V. B. The gut microbiome in health and in disease. *Curr. Opin. Gastroenterol.* **31**(1), 69 (2015).
5. Fan, Y. & Pedersen, O. Gut microbiota in human metabolic health and disease. *Nat. Rev. Microbiol.* **19**(1), 55–71 (2021).
6. Curtis, M. A., Diaz, P. I. & Van Dyke, T. E. The role of the microbiota in periodontal disease. *Periodontology 2000* **83**(1), 14–25 (2020).
7. Parente, E., Zotta, T. & Ricciardi, A. Microbial association networks in cheese: A meta-analysis. *bioRxiv.*https://doi.org/10.1101/2021.07.21.453196 *(2021).*
8. Chen, L. *et al.* Gut microbial co-abundance networks show specificity in inflammatory bowel disease and obesity. *Nat. Commun.* **11**(1), 1–12 (2020).
9. Faust, K. *et al.* Microbial co-occurrence relationships in the human microbiome. *PLoS Comput. Biol.* **8**(7), e1002606 (2012).
10. Faust, K. & Raes, J. Microbial interactions: From networks to models. *Nat. Rev. Microbiol.* **10**(8), 538–50 (2012).
11. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome datasets are compositional: And this is not optional. *Front. Microbiol.* **8**, 2224 (2017).
12. Weiss, S. *et al.* Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J.* **10**(7), 1669–81 (2016).
13. Jiang, D. *et al.* Microbiome multi-omics network analysis: Statistical considerations, limitations, and opportunities. *Front. Genet.* **10**, 995 (2019).
14. Tsilimigras, M. C. & Fodor, A. A. Compositional data analysis of the microbiome: Fundamentals, tools, and challenges. *Ann. Epidemiol.* **26**(5), 330–5 (2016).
15. Lloréns-Rico, V., Vieira-Silva, S., Gonçalves, P. J., Falony, G. & Raes, J. Benchmarking microbiome transformations favors experimental quantitative approaches to address compositionality and sampling depth biases. *Nat. Commun.* **12**(1), 1–12 (2021).
16. Röttjers, L. & Faust, K. From hairballs to hypotheses—Biological insights from microbial networks. *FEMS Microbiol. Rev.* **42**(6), 761–80 (2018).
17. Layeghifard, M., Hwang, D. M. & Guttman, D. S. Disentangling interactions in the microbiome: A network perspective. *Trends Microbiol.* **25**(3), 217–28 (2017).
18. McKnight, D. T. *et al.* Methods for normalizing microbiome data: An ecological perspective. *Methods Ecol. Evol.* **10**(3), 389–400 (2019).
19. McMurdie, P. J. & Holmes, S. Waste not, want not: Why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* **10**(4), e1003531 (2014).
20. Lovell, D., Pawlowsky-Glahn, V., Egozcue, J. J., Marguerat, S. & Bähler, J. Proportionality: A valid alternative to correlation for relative data. *PLoS Comput. Biol.* **11**(3), e1004075 (2015).

21. Matchado, M. S. *et al.* Network analysis methods for studying microbial communities: A mini review. *Comput. Struct. Biotechnol. J.* **19**, 2687–2698 (2021).
22. Jiang, S. *et al.* HARMONIES: A hybrid approach for microbiome networks inference via exploiting sparsity. *Front. Genet.* **11**, 445 (2020).
23. Connor, N., Barberán, A. & Clauset, A. Using null models to infer microbial co-occurrence networks. *PLoS One* **12**(5), e0176751 (2017).
24. Huson, L. W. Performance of some correlation coefficients when applied to zero-clustered data. *J. Mod. Appl. Stat. Methods* **6**(2), 17 (2007).
25. Kurtz, Z. D. *et al.* Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.* **11**(5), e1004226 (2015).
26. Tackmann, J., Rodrigues, J. F. M. & von Mering, C. Rapid inference of direct interactions in large-scale ecological networks from heterogeneous microbial sequencing data. *Cell Syst.* **9**, 286–296. https://doi.org/10.1016/j.cels.2019.08.002 (2019).
27. Menon, R., Ramanan, V. & Korolev, K. S. Interactions between species introduce spurious associations in microbiome studies. *PLoS Comput. Biol.* **14**(1), e1005939 (2018).
28. Friedman, J. & Alm, E. J. Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* **8**(9), e1002687 (2012).
29. Faust, K. & Raes, J. CoNet app: Inference of biological association networks using Cytoscape. *F1000Research* **5**, 1519 (2016).
30. Fang, H., Huang, C., Zhao, H. & Deng, M. CCLasso: Correlation inference for compositional data through Lasso. *Bioinformatics* **31**(19), 3172–80 (2015).
31. Hirano, H. & Takemoto, K. Difficulty in inferring microbial community structure based on co-occurrence network approaches. *BMC Bioinform.* **20**(1), 329 (2019).
32. Peschel, S., Müller, C. L., von Mutius, E., Boulesteix, A. L. & Depner, M. NetCoMi: Network construction and comparison for microbiome data in R. *Brief. Bioinform.* **22**(4), bbaa290 (2021).
33. Chen, L., He, Q., Wan, H., He, S. & Deng, M. Statistical computation methods for microbiome compositional data network inference. arXiv preprint arXiv:2109.01993 (2021).
34. Feng, K., Peng, X., Zhang, Z., Gu, S., He, Q., Shen, W. *et al.* iNAP: An integrated network analysis pipeline for microbiome studies. *iMeta*, e13 (2022).
35. Gupta, V. K. *et al.* A predictive index for health status using species-level gut microbiome profiling. *Nat. Commun.* **11**(1), 1–16 (2020).
36. Simon, H. Y., Siddle, K. J., Park, D. J. & Sabeti, P. C. Benchmarking metagenomics tools for taxonomic classification. *Cell* **178**(4), 779–94 (2019).
37. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**(15), 2114–20 (2014).
38. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**(4), 357–9 (2012).
39. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**(1), 1–13 (2019).
40. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: Estimating species abundance in metagenomics data. *PeerJ Comput. Sci.* **3**, e104 (2017).
41. Cao, Q. *et al.* Effects of rare microbiome taxa filtering on statistical analysis. *Front. Microbiol.* **11**, 3203 (2021).
42. Meinshausen, N. & Bühlmann, P. High-dimensional graphs and variable selection with the lasso. *Ann. Stat.* **34**(3), 1436–62 (2006).
43. Bastian, M., Heymann, S. & Jacomy, M. Gephi: An open source software for exploring and manipulating networks. In: *Third International AAAI Conference on Weblogs and Social Media* (2009).
44. Traag, V. A., Waltman, L. & Van Eck, N. J. From Louvain to Leiden: Guaranteeing well-connected communities. *Sci. Rep.* **9**(1), 1–12 (2019).
45. Blondel, V. D., Guillaume, J. L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**(10), P10008 (2008).
46. Fujio-Vejar, S. *et al.* The gut microbiota of healthy Chilean subjects reveals a high abundance of the phylum Verrucomicrobia. *Front. Microbiol.* **8**, 1221 (2017).
47. Rinninella, E. *et al.* What is the healthy gut microbiota composition? A changing ecosystem across age, environment, diet, and diseases. *Microorganisms* **7**(1), 14 (2019).
48. Consortium, H. M. P. *et al.* Structure function and diversity of the healthy human microbiome. *Nature* **486**(7402), 207 (2012).
49. Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A. & Alm, E. J. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nat. Commun.* **8**(1), 1–10 (2017).
50. Mosca, A., Leclerc, M. & Hugot, J. P. Gut microbiota diversity and human diseases: Should we reintroduce key predators in our ecosystem?. *Front. Microbiol.* **7**, 455 (2016).
51. Rizzatti, G., Lopetuso, L., Gibiino, G., Binda, C. & Gasbarrini, A. Proteobacteria: A common factor in human diseases. *BioMed Res. Int.* **2017**, 9351507 (2017).
52. Shin, N. R., Whon, T. W. & Bae, J. W. Proteobacteria: Microbial signature of dysbiosis in gut microbiota. *Trends Biotechnol.* **33**(9), 496–503 (2015).
53. Banerjee, S., Schlaeppi, K. & van der Heijden, M. G. Keystone taxa as drivers of microbiome structure and functioning. *Nat. Rev. Microbiol.* **16**(9), 567–76 (2018).
54. Jousset, A. *et al.* Where less may be more: How the rare biosphere pulls ecosystems strings. *ISME J.* **11**(4), 853–62 (2017).
55. Lynch, M. D. & Neufeld, J. D. Ecology and exploration of the rare biosphere. *Nat. Rev. Microbiol.* **13**(4), 217–29 (2015).
56. Jacomy, M., Venturini, T., Heymann, S. & Bastian, M. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS One* **9**(6), e98679 (2014).
57. Milanese, A. *et al.* Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat. Commun.* **10**(1), 1–11 (2019).
58. Lloyd-Price, J. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**(7758), 655–62 (2019).

Acknowledgements

The authors thank Dr. Yong Yeol Ahn for helpful discussions of network based analyses.

Author contributions

T.L. and Y.Y. conceived the research. T.L and Y.Y. wrote programs to process data. T.L. processed the data and preformed all analysis. T.L. and Y.Y. interpreted results and participated in preparing the manuscript.

Funding

The NIH grant R01AI143254 and NSF grant EF-2025451.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-22541-1.

Correspondence and requests for materials should be addressed to Y.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.