



OPEN

Bacterial associations in the healthy human gut microbiome across populations

Mark Loftus^{1,3}, Sayf Al-Deen Hassouneh^{1,3} & Shibu Yooseph²✉

In a microbial community, associations between constituent members play an important role in determining the overall structure and function of the community. The human gut microbiome is believed to play an integral role in host health and disease. To understand the nature of bacterial associations at the species level in healthy human gut microbiomes, we analyzed previously published collections of whole-genome shotgun sequence data, totaling over 1.6 Tbp, generated from 606 fecal samples obtained from four different healthy human populations. Using a Random Forest Classifier, we identified 202 signature bacterial species that were prevalent in these populations and whose relative abundances could be used to accurately distinguish between the populations. Bacterial association networks were constructed with these signature species using an approach based on the graphical lasso. Network analysis revealed conserved bacterial associations across populations and a dominance of positive associations over negative associations, with this dominance being driven by associations between species that are closely related either taxonomically or functionally. Bacterial species that form network modules, and species that constitute hubs and bottlenecks, were also identified. Functional analysis using protein families suggests that much of the taxonomic variation across human populations does not foment substantial functional or structural differences.

The community of microbial cells in the human gut is estimated to be comparable in magnitude to the number of human cells¹. This community, deemed the human gut microbiome, is mainly composed of bacteria, archaea, fungi, and viruses, with bacteria being the largest constituent. These bacterial cells exist in a complex consortium of ecological and metabolic interactions that ultimately influence the taxonomic and functional profile of the microbial community, as well host health. The gut microbiome of healthy individuals is believed to be mainly symbiotic and is known to play important roles in host metabolism, immunological modulation and development, cell signaling, pathogen colonization resistance, and mucosal regeneration and homeostasis^{2–4}.

The continued stability of this community and its functions, i.e. homeostasis^{5,6}, is important and its disruption, broadly described as ‘dysbiosis’⁷, has been associated with numerous diseases including, but not limited to: diabetes⁸, cardiovascular disease^{9,10}, obesity¹¹, inflammatory bowel disease^{12,13}, and various cancers¹⁴. However, it remains unclear whether disease onset is the consequence or cause of the microbiome community disruption. Furthermore, what constitutes a healthy gut microbiome is still under investigation due to the overwhelming amount of bacterial species found in the gut, and the large variation in their carriage rates across human populations and individuals^{15,16}. These issues are of great importance as one of the ultimate goals of microbiome research is to modulate the community from a ‘dysbiotic’ state into a healthy ‘homeostatic’ one.

Early research towards this goal chose to limit their focus to taxonomic differences between healthy and disease microbiomes^{17–19}. While these comparisons are valuable, since the bacterial community taxonomic profile generally represents the potential metabolic and transcriptional profiles that are present within the ecosystem; simply profiling the community fails to acknowledge the underlying bacterial associations and the impact they exert on both the microbial ecosystem and host health. In fact, many studies within natural systems and animal hosts have shown that the associations (positive and negative) between bacteria are an important foundation for the continued stability and proper functioning of these ecosystems^{20–25}. As such, it is of great importance to assess the relationships that exist between bacteria within the healthy human gut microbiome in order to better understand the ecological associations important for the structure and maintenance of the gut microbiome and its related processes. Naturally, this raises an important question: are there similarities in the structural features

¹Burnett School of Biomedical Sciences, Genomics and Bioinformatics Cluster, University of Central Florida, Orlando 32787, USA. ²Department of Computer Science, Genomics and Bioinformatics Cluster, University of Central Florida, Orlando, FL 32816-2993, USA. ³These authors contributed equally: Mark Loftus and Sayf Al-Deen Hassouneh. ✉email: Shibu.Yooseph@ucf.edu

of bacterial association networks in human gut microbiomes across healthy populations, and if so, are there conserved associations?

Microbial associations in a community are characterized by both direct and indirect interactions between the constituents²⁶. In this paper, we depict these associations using a weighted graph (network) in which the nodes represent bacterial species and an edge between two nodes represents an association between the corresponding species, with the edge weight capturing the strength of the association. This framework enables us to model both positive and negative associations between species, and thus can help to shed light on cooperation and competition between species in the community. Once a network is constructed, an analysis of the various topological properties of the network can enable us to decipher the underlying ecological rules associated with the microbial ecosystem. These networks also provide the ability to determine the relative importance of species for ecosystem structure and function.

Microbial association networks are typically constructed from a sample-taxa count matrix generated by collecting multiple samples from the community and determining the taxa counts in each sample. With the availability of high-throughput and low-cost DNA sequencing technologies, these counts are generated by sequencing the collected biological samples. Microbiome sequence data are generated either using a targeted approach, involving the sequencing of a taxonomic marker gene (e.g., the 16S ribosomal RNA gene)²⁷ or using a whole-genome shotgun (WGS) sequencing approach²⁸. However, estimates of taxa abundances using 16S rRNA sequences can be confounded by several factors including the presence of multiple copies and variants of the 16S rRNA gene in genomes, and the lack of taxonomic resolution in the selected variable region of the 16S gene^{29,30}. Conversely, WGS data can be used to provide more accurate estimates of genome relative abundances as well as higher resolution taxonomic classification, compared to 16S rRNA data^{31,32}. Regardless of sequencing approach, the taxa count data generated by DNA sequencing are compositional in nature and provide only relative abundance information of the constituent taxa³³. This poses challenges for inferring associations, and the computation of measures like correlation directly from the observed sequence counts can be misleading³⁴. While several methods have been proposed for constructing association networks that address this challenge³⁵, here we use a Gaussian Graphical Model (GGM) framework on Centered Log-Ratio (CLR) transformed count data to construct an association network^{36,37}.

We are motivated by the observation that the covariance matrix of a multivariate Gaussian distribution used to fit log-transformed *relative* count data provides a good approximation to the covariance matrix of the log-transformed *absolute* count data^{36,37}. The GGM framework also enables the modeling of conditional dependencies of the random variables that represent taxa abundances. The adjacency matrix of the association network that we construct is the *inverse* covariance matrix (i.e. the *precision* matrix) of the underlying multivariate Gaussian distribution used in the GGM. This graph has the property that an edge exists between two nodes if and only if the corresponding entry in the precision matrix is non-zero. A zero entry in the precision matrix indicates conditional independence between the two corresponding random variables. We also incorporate sparsity in our framework using the *l1*-penalty norm and construct sparse association networks using the graphical lasso method (glasso)³⁸.

In this study we investigate bacterial association networks in gut microbiomes across four healthy human populations. Previous studies analyzing bacterial association networks have mainly used 16S rRNA data, and given its lower taxonomic resolution, these studies have analyzed associations at the genus level³⁹. Instead, here we use a large collection of WGS samples from multiple human populations to investigate bacterial associations at the species level. We use a machine learning algorithm to identify a set of signature species that can accurately distinguish between the different healthy populations. Using these signature species, we construct networks by employing a glasso method that incorporates a bootstrapping⁴⁰ approach to reduce the number of false positive edges inferred⁴¹. We analyze these networks to assess the theoretical ecology, and potential importance of species within healthy human gut microbial communities.

Results

Signature species in the healthy human gut microbiome. For each cohort, the prevalence of individual species across all samples was measured and plotted. All cohorts exhibited a skewed bi-modal distribution (Fig. 1a). The first peak in the distribution was centered around a prevalence of 10%, while the second peak occurred around a prevalence of 90%. This skewed bi-modal distribution has been previously observed in a microbial community, and organisms that were highly prevalent were deemed the ‘abundant core’ as they were found to account for the majority of total sample abundances⁴². The 90% prevalent species set for each cohort consisted of 127 (American), 109 (Indian), 182 (European), and 146 (Japanese) species respectively, and these species were found to account for a large majority of the total sample proportions, the median values for the cohorts were 0.93 (American), 0.93 (Indian), 0.87 (European), and 0.81 (Japanese) (Fig. 1b). We utilized a Random Forest Classifier (RFC) framework to determine the effect of prevalence threshold values on the ability to distinguish between cohorts using the taxonomic profiles of the constituent samples. For each prevalence threshold value, a single input feature set was generated to construct the classifier; this feature set was produced by taking the union of the bacterial species sets for the four cohorts (at that prevalence threshold value). The RFC was able to distinguish between cohorts with an F1-score > 0.85 for all tested prevalence thresholds (0%, 20%, 40%, 50%, 60%, 80%, 90%, 100%), but demonstrated the highest F1-score at the 90% threshold, even though less than 10% of the original species remained (Supplemental 1). Based on this analysis, we define the set of *signature species* to be the union of the prevalent (> 90%) species sets from the four cohorts. The signature species set consisted of 202 species and was used for constructing the bacterial association network for each cohort. We explored the variability in signature species relative abundance between samples using principal components analysis (PCA) applied to the CLR-transformed data (Fig. 1c). PCA showed evidence for separation of samples

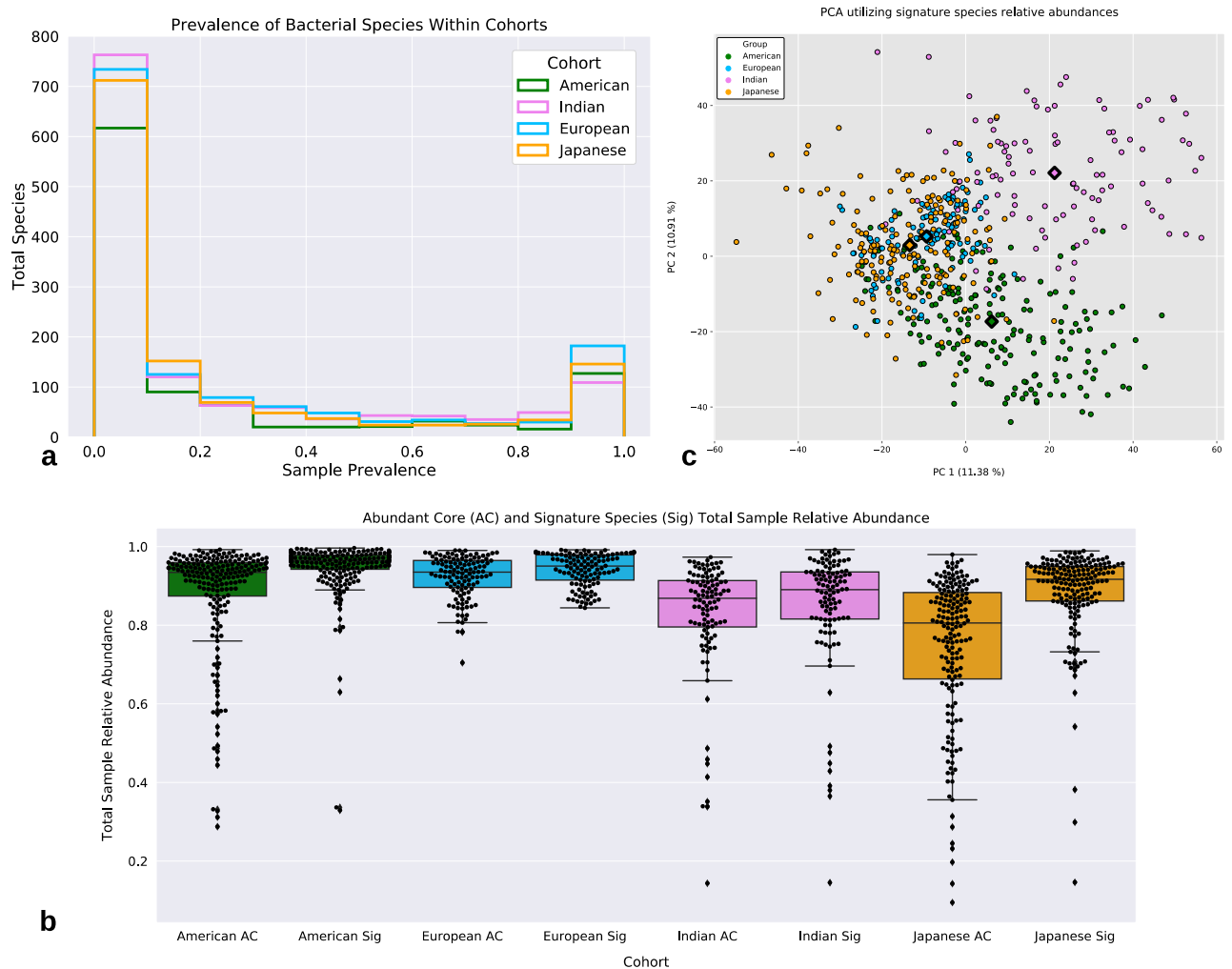


Figure 1. ‘Abundant cores’ and Signature Species. **(a)** All cohorts exhibit a bimodal distribution for species prevalence. Species that are prevalent in 90% or more samples within a cohort is considered a member of that cohort’s ‘abundant core.’ **(b)** The proportion of total sample relative abundance each cohort’s ‘abundant core’ species and the union of all ‘abundant cores’ species (i.e., Signature Species/Sig). The ‘abundant core’ microbiota is shown to account for the bulk of reads mapped within each sample. Each dot represents a sample from that cohort. **(c)** PCA demonstrating the lack of distinct clustering of samples from different cohorts based on the CLR-transformed relative abundance data of the signature species. Samples from the Indian and American cohorts appear to separate from the rest of the cohorts however, samples from the other two cohorts demonstrate little separation. The diamonds indicate cluster centroids.

from the Indian and American cohorts, but ultimately the PCA only explained a small amount of the total variance (PC1: 11.38%, PC2: 10.91%).

Bacterial association networks. Prior to its application on the cohort data, the network inference method with bootstrapping was tested on synthetic data (see Supplemental) notably, most graph-types were inferred with an F1-score above 0.7 (band: 0.974, hub: 0.885, random: 711, cluster: 0.692, scale-free: 0.416) (Supplemental 2a). Furthermore, we demonstrate that as the sample-to-taxa ratio increases, F1-scores approach 1, and all groups demonstrate mean F1-scores above 0.9 (Supplemental 2a). Finally, we observe that our network inference method tends to underestimate edge weights, and on average the estimated edge weights are 53.23% of the actual edge weights (Supplemental 2b). A bacterial association network was constructed for each cohort using the CLR-transformed relative abundances of the signature species (see “Methods”). Each network was modeled as an undirected graph consisting of nodes and edges (Fig. 2). At a high-level, differences in the structure of the four networks were apparent. The European, Japanese, and Indian networks exhibited a high density of edges occurring between nodes from the phylum *Firmicutes*, whereas the American network had the largest density of edges existing between nodes from the phylum *Bacteroidetes*. Positive associations were dominant in all networks (American: 0.98, Indian: 0.97, European: 0.96, Japanese: 0.96), and negative associations involve nodes from the phylum *Firmicutes*. Network topology was studied by calculating the following network properties: average shortest path length (ASPL), transitivity, modularity, degree assortativity, and genera assortativity

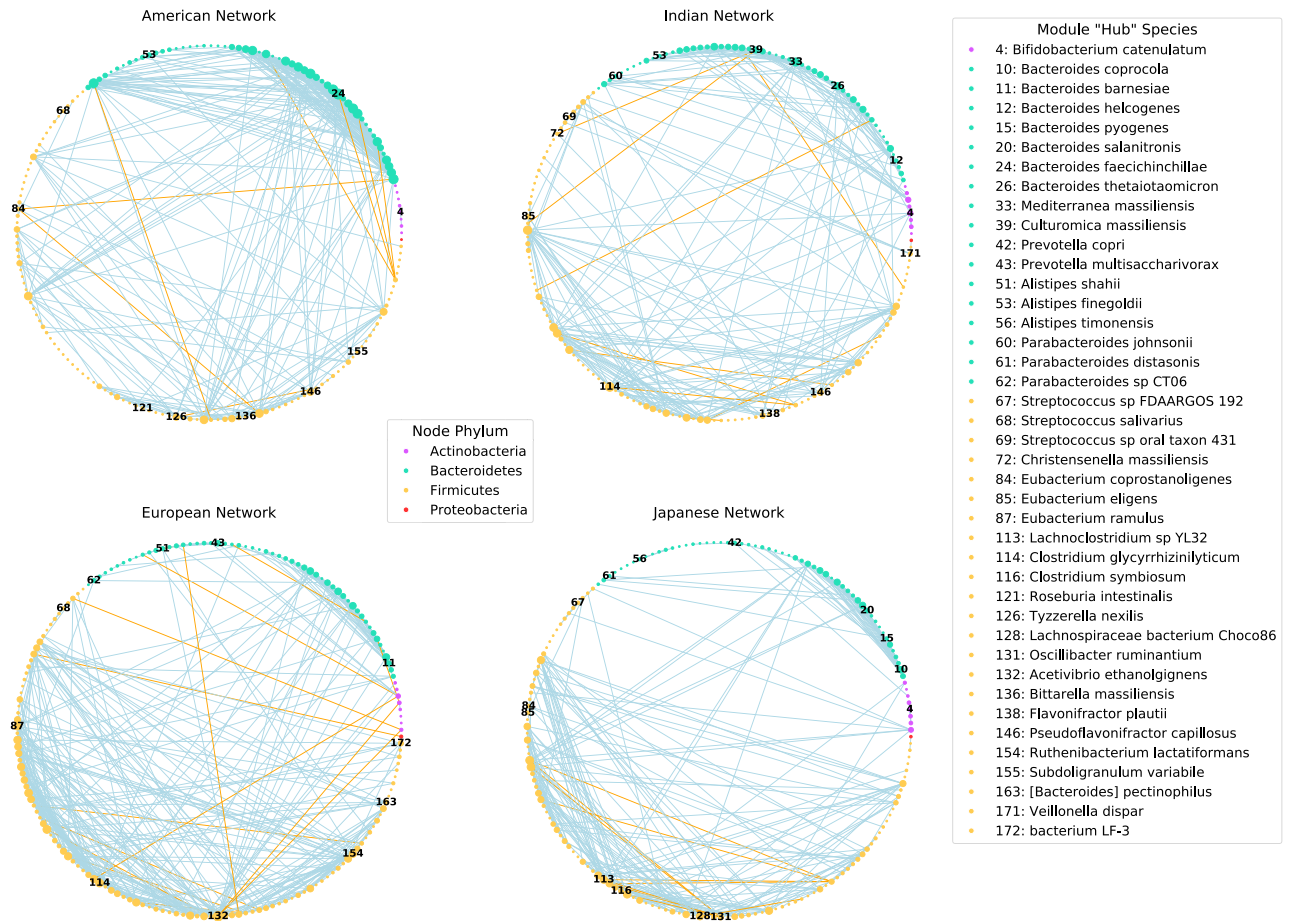


Figure 2. Species-level bacterial association networks. Network modeling of associations between (173/202) signature species within each network. A total of 29 species were not shown as they had zero edges in all networks. Node color designates the phylum each species belongs to, node size is reflective of node degree, and edge color represents if the association is positive (blue) or negative (orange). Nodes are ordered counterclockwise around the circle by the alphabetical order of the concatenated string of all taxonomic levels. Nodes that are numbered correspond to species with the highest degree centrality within modules, designated as “hubs”. Brackets around [*Bacteroides*] *pectinophilus* indicate that it is misclassified (i.e., placed incorrectly in a higher taxonomic rank and awaiting to be formally renamed). We utilized Blast to designate [*Bacteroides*] *pectinophilus* as belonging to the phylum *Firmicutes*⁹¹. For a full list of species shown and not shown within network models see Supplemental.

Network	Nodes	Edges	Density	ASPL	Transitivity	Modularity	Degree assortativity	Genera assortativity
American	202	338	0.017	1.539 (–, ***)	0.487 (+, ***)	0.475 (+, *)	0.338 (+, ***)	0.144 (+, ***)
Indian	202	273	0.013	1.874 (–, *)	0.452 (+, ***)	0.667 (+, ***)	0.330 (+, ***)	0.163 (+, ***)
European	202	386	0.019	1.369 (–, ***)	0.353 (+, ***)	0.681 (+, ***)	0.158 (+, *)	0.196 (+, ***)
Japanese	202	274	0.013	1.444 (–, ***)	0.471 (+, ***)	0.755 (+, ***)	0.308 (+, ***)	0.242 (+, ***)

Table 1. Cohort network topological properties. Network topological properties calculated for each cohort’s network. The plus (+) or minus (–) sign indicates that the network property was greater or lower than the average of 1000 random networks. Stars indicate that the network property was statistically significantly different (P-value: * < 0.05, ** < 0.01, *** < 0.001) based on the Monte Carlo simulations.

(see “Methods”) (Table 1). These properties were compared to random networks using Monte Carlo simulations (see Supplemental). All cohort networks were deemed non-random in their topology and exhibited significantly low ASPL (all P-values < 0.05), significantly high modularity (all P-values < 0.01), significantly high transitivity (all P-values < 0.001), significantly high genera assortativity (all P-values < 0.001) and significantly high degree assortativity (all P-values < 0.01), relative to the random networks. The low ASPL within networks suggest that nodes are connected to one another through short paths within the network. The high transitivity and modularity indicate that nodes form cliques and networks exhibit compartmentalization (modules), respectively. Lastly,

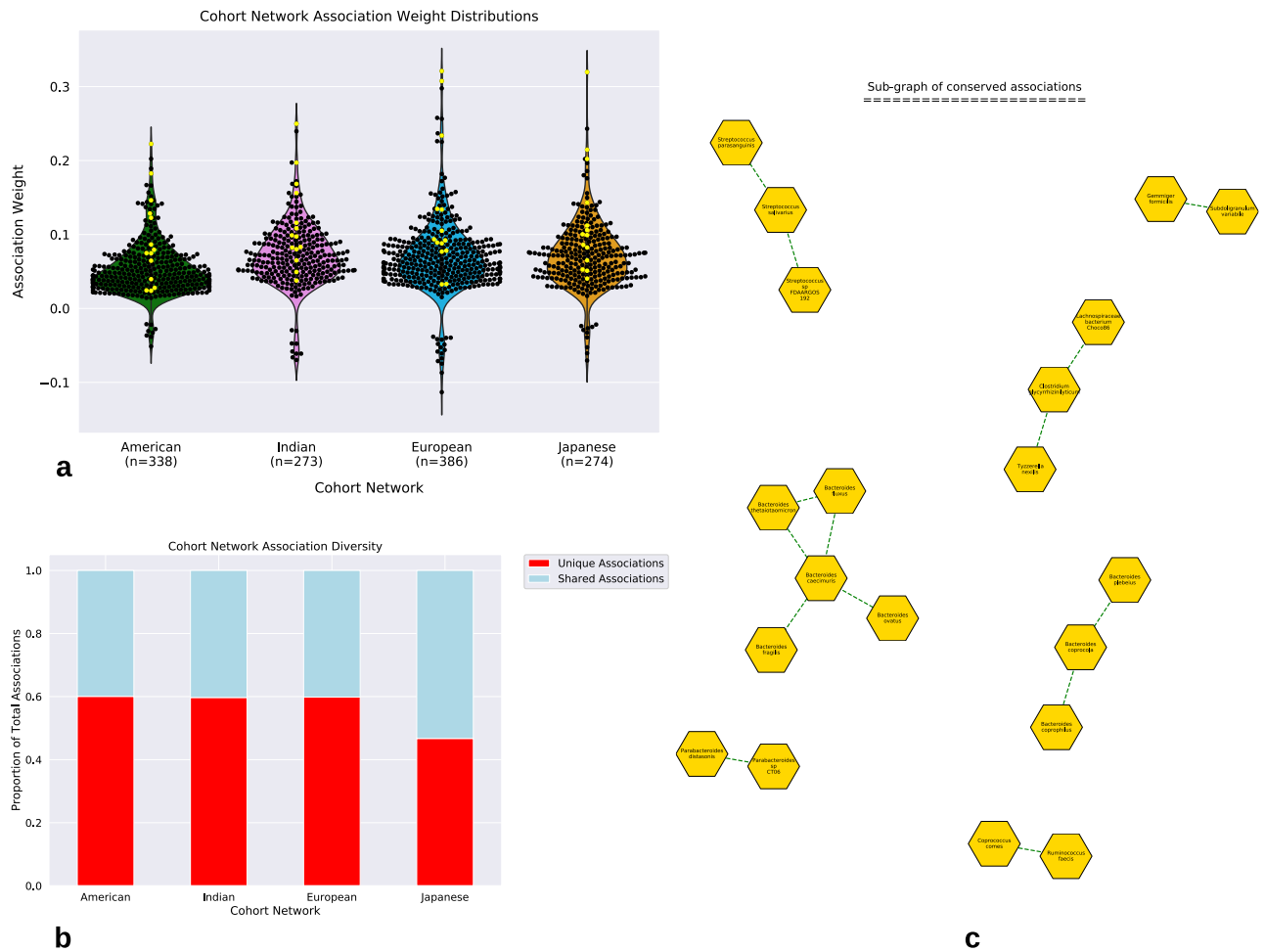


Figure 3. Cohort network association analysis. **(a)** The distribution of bacterial association weights within each cohort's network, dots (black and yellow) and (n) represent total associations. Yellow dots represent species associations that were found shared across all networks. **(b)** The proportion of associations within each cohort's network that are unique (red) or shared (blue) with at least one other network. **(c)** Sub-graph displaying only the 20 conserved nodes (species) and 14 edges (associations) retained across all cohorts.

the high (assortative) degree assortativity and genera assortativity portrays that nodes tend to form connections to other nodes that have a similar degree and taxonomy.

Theoretical ecology based on bacterial association networks. All cohort networks were found to contain highly similar distributions of association (edge) weights, where positive associations were more frequent and greater in magnitude than negative associations (Fig. 3a). Furthermore, a large percentage of associations (American: 40%, Indian: 40%, European: 40%, Japanese: 53%) were found to be shared with at least one other network, however, the Japanese network shared the largest proportion of associations with all other networks (American: 26%, Indian: 22%, European: 33%) and the Indian network the least with all other networks (American: 18%, Japanese: 22%, European: 16%) (Supplemental 3). Interestingly all shared associations were positive (Fig. 3b). A conserved structure of 14 associations, composed of 20 species (Fig. 3c), mainly from the genus *Bacteroides*, was observed to be contained within all networks (Supplemental 4). Many of these conserved associations were associations with relatively higher edge weights (Fig. 3a). No negative association was retained across networks. However, viewed at the higher taxonomic rank for those species involved in negative associations, we observed that across all cohort networks, members from the phylum *Firmicutes* were involved in a large percentage of the negative associations (American: 100%, Indian: 100%, European: 62.5%, Japanese: 100%), and specifically these negative associations were mainly occurring between species from the order *Clostridiales* (American: 25%, Indian: 89%, European: 56%, Japanese: 100%) (Supplemental 5). We next explored the taxonomic relationship between species and their association type (positive or negative) (Fig. 4a), as well as the genome functional profile dissimilarities, according to Bray–Curtis dissimilarity, between network neighbors against their association weight (Fig. 4b). We found that most positive associations take place between bacteria that are more taxonomically and functionally similar, while negative associations were never found between species within the same genus, or between species with low genome functional profile distance (<0.2), and linear regression showed a negative correlation (p -value <0.05) between association weight and partner genome functional distance (Supplemental 6).

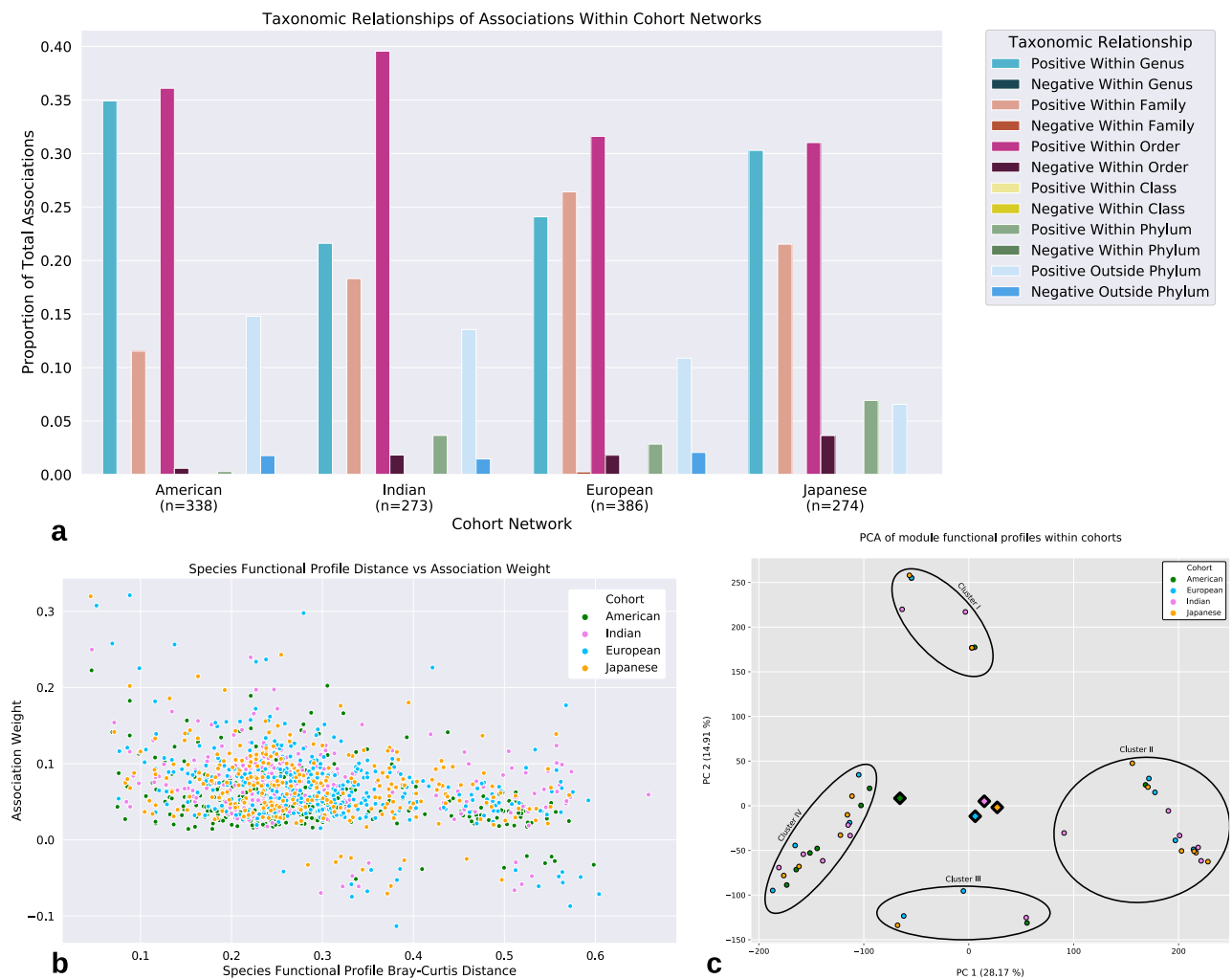


Figure 4. Taxonomic and functional relationships between species. **(a)** Proportion of associations within each cohort's network that are either positive or negative at the lowest level of taxonomic relation (n = total associations). Most positive associations appear between taxonomically similar species. **(b)** Association weight vs Bray–Curtis distance of genome functional profiles between network partners. Positive associations between functionally similar species are both common and greater in strength than negative associations. There appears to be a minimal distance between genome functional profiles before a negative association is demonstrated. **(c)** An asynchronous LPA was used to analyze the modules composing the networks of each cohort. Each dot represents the aggregated TIGRFAM profiles of an individual module found by aLPA and the diamonds represent the cohort centroids. Four distinct clusters were found, and each cohort was represented within each cluster. The American cohort appears to be biased towards Cluster IV, however the other cohorts do not appear overtly biased to any one cluster.

Network cliques and module detection. As our networks exhibited both high transitivity and modularity, we sought to investigate the cliques and modules of species contained within them. We first found all cliques of three species (1588 unique cliques) within our networks (see “Methods”). These triadic cliques are important to understand because their formation provides stability to the community structures existing between species^{43,44}. Of these cliques: 113 were shared in at least 1 other network, 8 were shared across three networks, and only 1 (*Bacteroides caecimuris*, *Bacteroides fluxus*, *Bacteroides thetaiotaomicron*) was found in all networks. In total, 66 genera were shown to participate in clique formation, however, cliques were shown to be mainly (American: 29%, Indian: 72%, European: 64%, Japanese 55%) formed between species from differing genera (Supplemental 7a). Species from the genus *Bacteroides* were found to be involved in the largest percentage of cliques (American: 21.0%, Indian: 4.0%, European: 4.9%, Japanese: 5.8%) within most cohort networks (Supplemental 7b). Interestingly, the cliques that contained species from *Bacteroides* were also the most retained (American: 20.9%, Indian: 8.5%, European: 8.5%, Japanese: 10.8%) across all cohorts (Supplemental 7c).

Following clique analysis, we performed module detection utilizing an asynchronous Label Propagation Algorithm (aLPA) (see Supplemental) which identified a total of 49 modules (American: 10, European: 11, Indian: 14, Japanese: 14) that contained 3 or more members⁴⁵ (Supplemental 8). The quality of network partitioning by the module detection algorithm (performance) was analyzed (American: 0.96, Indian: 0.98, European: 0.94,

Japanese: 0.98) showing that the majority of edges between nodes were contained within modules (see Supplemental). PCA was utilized to examine the variance between Module Functional Profiles (MFP's) of the different cohort (Fig. 4c). This analysis revealed MFPs fell within one of four clusters, and each cohort had representation within each cluster. Taxonomic and functional characteristics of the clusters were analyzed. Cluster I contained modules formed mainly by the genera *Streptococcus* and *Bifidobacterium* (Fig. 5a). Cluster II modules were mainly composed of species from the genera *Alistipes*, *Bacteroides*, and *Prevotella* (Fig. 5b). Cluster III modules were dominated by the genera *Bacteroides* (Fig. 5c). Cluster IV modules were mainly composed of species from the genera *Blautia*, *Eubacterium*, *Lachnospirillum*, and *Ruminococcus* (Fig. 5d). Functional analysis of clusters revealed unique roles in each cluster. Cluster I displayed an increase in roles linked to toxin production, protein secretion, anaerobic metabolism, nucleic acid metabolism, and a decrease in roles linked to thiamine biosynthesis. Cluster II displayed an increase in roles linked to cellular metabolism and protein degradation, displayed a decrease in roles linked to cell division and signal transduction. Cluster III displayed an increase in roles linked to chemoautotrophy, sulfur and phosphorous metabolism, and DNA metabolism. Lastly, cluster IV displayed an increase in roles tied to transcription factors and a decrease in roles associated with adaptation to atypical conditions (Supplemental 9).

We next analyzed the sample functional profiles using PCA (Supplemental 10a). PCA explained a modest amount of variance (PC1: 27.82%; PC2: 5.99%) although samples between cohorts were found to overlap. When analyzing the Cohort Functional Role Profiles (CFRP's), only 11 differences, when comparing the signs (+/-), out of the 113 found roles were found, and only the European cohort exhibited more than two differences (Supplemental 10b).

Node centrality analysis. We utilized degree and betweenness centrality measurements to identify “hub” and “bottleneck” nodes, respectively, within our networks (see Supplemental). These centrality measurements were selected because ‘hubs’ and ‘bottlenecks’ are nodes that could have strong influence within a network and have been utilized previously to identify important species within microbial ecosystems^{21,23,46}. Considering all cohort networks were deemed assortative in respect to their degree assortativity, we did not expect to find network “hub” nodes. However, we did find that nearly all modules, within each cohort, were disassortative in their degree assortativity which hinted at “hub” nodes existing within modules (Supplemental 11). For these reasons, we chose to select the node within each module that exhibited the highest degree (see Fig. 2), and the top 10 nodes within each network with the highest betweenness. Across all cohorts we found variation in the species deemed module ‘hubs’ and ‘bottlenecks’ (Supplemental 12a), although at the genus level there was a large amount of agreement (Supplemental 12b). In at least three out of the four cohorts, species from *Bacteroides*, *Alistipes*, *Bifidobacterium*, *Eubacterium*, *Parabacteroides*, and *Streptococcus* were designated as ‘hubs’, whereas species from *Bacteroides* and *Lachnospirillum* were designated as ‘bottlenecks’.

Discussion

In this study, we used WGS data in conjunction with a network inference method that is robust to sequence data compositionality to analyze the associations occurring between species within the healthy human gut microbiome across different populations. The association networks were constructed utilizing the signature species.

We demonstrated that bacterial association networks, across all cohorts, do not have the same properties as random networks. However, relative to each other, the networks of the four cohorts display similar properties. Random networks are known to contain short average path lengths, low node clustering, and high modularity^{46,47}. Compared to random networks each cohort network was found to exhibit significantly shorter average shortest path lengths, significantly higher transitivity (clustering), significantly higher modularity, significantly higher degree assortativity, and significantly higher genera assortativity. We posit that the similarities in network properties reflect an organization of the bacterial community that is important to underlying ecological processes. For instance, the short average path lengths within our networks could imply rapid signaling between bacterial species, potentially facilitating swift changes in community metabolism. This is supported by previous studies demonstrating that the human gut microbiome exhibits rapid alterations in bacterial metabolism and abundance in conjunction with change in host diet¹⁹.

In addition to exhibiting similar properties, cohort networks also shared a large percentage of associations (American: 40%, Indian: 40%, European: 40%, Japanese: 53%), including a conserved set of 14 positive associations composed of 20 species. These conserved associations may be indicative of strong partner fidelity, important ecological relationships, or potentially obligate partnerships. Furthermore, we found that taxonomically and functionally similar species tended to have positive associations. This finding was unexpected as some previous studies on microbial ecosystems, including the human gut^{48–50}, have shown negative interactions between bacteria (competition, predation, etc.) should be the dominant form of interaction⁵¹, especially when those bacteria are taxonomically or functionally alike⁵². The differences between our results and the aforementioned research may be due to their use of non-transformed data and pairwise analysis. It has been noted that compositional data exhibit a negative correlation bias³³, and thus, failing to account for the compositional nature of sequencing data may lead to inferring more negative associations than those that actually exist. In fact, a previous comparison of compositionally robust network methodologies demonstrated that the majority of associations for these methodologies are positive³⁷. Our findings would suggest that kin-selection⁵³ (positively associating with those of similar lineage to directly or indirectly pass on one's genes), as opposed to competitive exclusion⁵⁴ (bacteria with similar lineage or functionality are more likely to compete within a habitat), is more prevalent within the healthy gut microbiome. This observation cannot be excluded as there is precedence within microbial ecosystems for the co-occurrence of bacteria with similar genetic traits^{52,55}, and studies on bacterial dynamics in the gut that

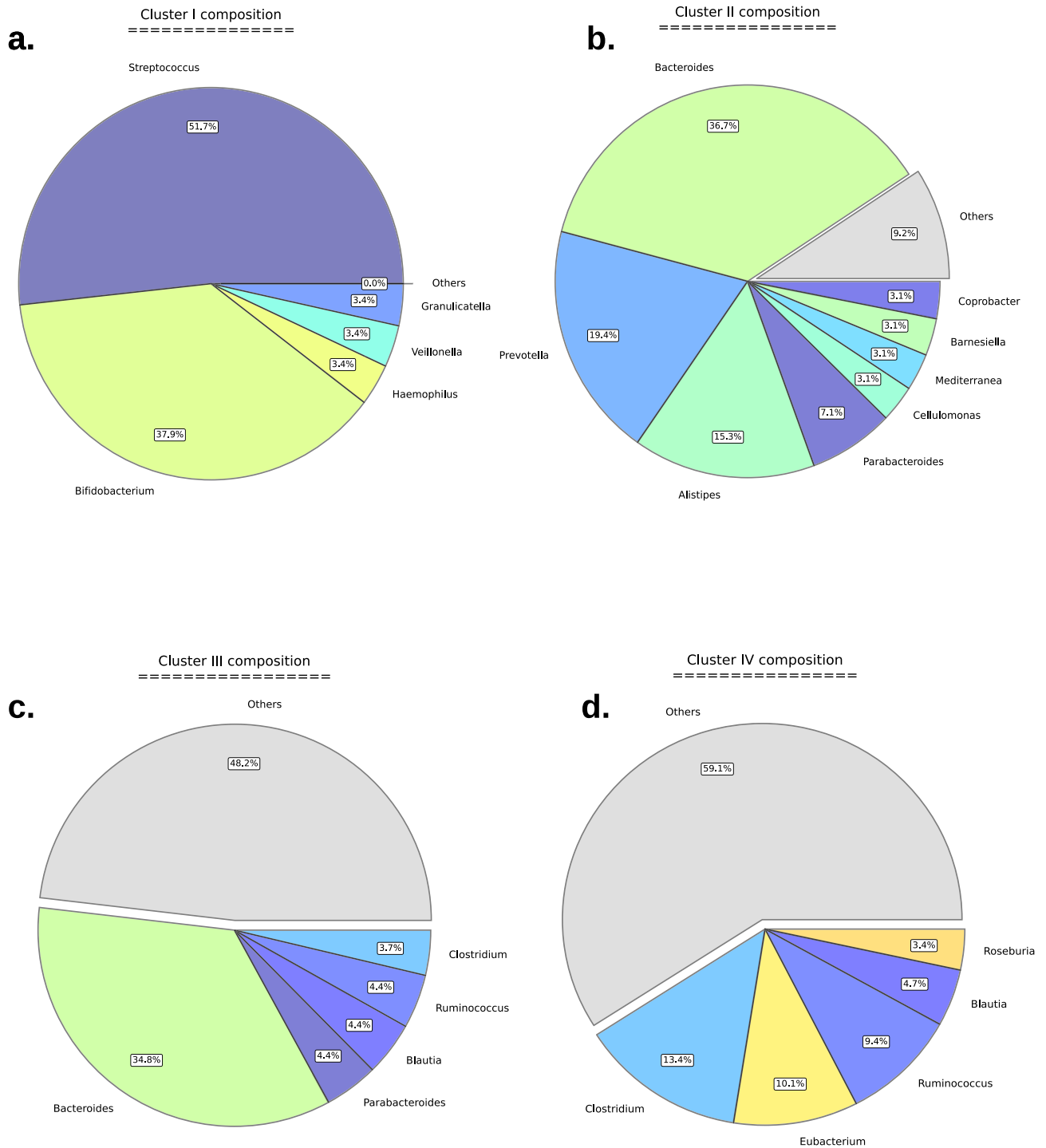


Figure 5. Pie plots of the cluster taxonomy. Pie plots demonstrating genus-level taxonomic compositions within each of the module clusters. Clusters were determined using PCA of module functional profiles for each module. (a) Cluster I is dominated by members of the *Streptococcus* and *Bifidobacterium* genera and no genus represents less than 3% relative abundance. (b) Members of the *Bacteroides* genus are also the most abundant in the Cluster II, however the *Prevotella* and *Allistipes* genera are also abundant and account for > 70% of abundance when combined with *Bacteroides*. There are 6 genera with relative abundances below 3%. (c) Members of the *Bacteroides* genus are the most abundant in the Cluster III and there are 49 genera with relative abundances below 3%. (d) There are only 5 genera above 3% relative abundance and 44 genera below 3% with no one genus showing greater than 15% relative abundance. Genera with < 3% relative abundance were placed in the ‘Others’ category.

suggest close relatives to bacteria currently present in the gut are more likely to be recruited into the community, i.e., phylogenetic under-dispersion (nepotism) hypothesis⁵⁶.

Within all cohorts, positive associations were not only the most dominant form of association, but also the only associations that were shared across networks. This finding seems logical as within the anoxic environment of the gut, bacterial energy production is limited which would make positive associations, such as mutual cross-feeding, preferable in order to produce and utilize energy more efficiently⁵⁷. In addition, ecological community theory suggests that partitioning of resources in space and time drive coexistence⁵⁸, and bacteria within the human gut microbiota are known to exhibit diurnal fluctuations⁵⁹ and exist in distinct spatial organizations^{60–62}. Furthermore, positive associations between species are also known to alleviate ecosystem stresses and allow for a greater diversity of organisms to coexist⁶³, and the healthy gut microbiome has a high level of biodiversity⁶⁴. However, it is important to be cognizant that a positive association between species does not rule out the presence of a negative interaction completely, as negative interactions between species can still have a net positive result if an increased survival rate is occurring, as well as to understand that these positive associations are not always indicative of cooperative activities as they could simply reflect a common preferred environmental niche⁶³. In contrast to the large proportion of shared positive associations, negative associations were always unique to a specific cohort; however, as we viewed the higher-level taxonomic ranking of species involved in negative associations, we found that across all cohorts most negative associations were occurring between species from the order *Clostridiales*. Species from the order *Clostridiales* are known to be largely cellulolytic, in that they mainly hydrolyze the polysaccharide cellulose⁶⁵. This limited nutritional niche could theoretically create competition between *Clostridiales* sp., and in any case, these associations might be important for community stability as negative associations within microbial communities are thought to be an important stabilizing force⁵⁰. In our analysis, 29 (out of the 202) species were found to exhibit no associations (positive or negative) across all networks. It may be possible that these species have very low strengths of association with some of the other species (i.e., partial correlation values below the detection threshold). It is also possible that some of these species occupy a unique metabolic niche in which they are capable of utilizing a specific resource for their metabolic requirement that is inaccessible to other microorganisms thereby limiting any cooperative or competitive actions.

While the healthy human gut microbiome is indeed routinely described as stable⁶⁴, the low abundance of negative associations within our networks suggests that the gut microbiome would be more vulnerable to positive feedback loops between species which could result in instability⁵⁰. We hypothesize that the high modularity found within all cohort networks could mitigate the vulnerability to positive feedback loops as high network modularity has been shown to have a stabilizing effect⁴⁷. We used a module detection algorithm to identify groups of highly connected species within our networks. The algorithm identifies modules of species which have previously been noted to benefit by growing together (e.g., *Bifidobacterium* sp.)⁶⁶. As we analyzed the variance between module functional profiles, using PCA, we found that modules gravitated towards one of four clusters. Although some cohorts had a greater proportion of modules within certain clusters, all cohorts had some level of representation within each cluster. Upon further analysis, we were able to find distinct functional and taxonomic differences between module clusters, but we were not able to distinguish overt functional differences between CFRP's. This implies that a general set of functions is present in each healthy population regardless of taxonomic differences. These module clusters may be indicative of niches that are retained in the healthy human gut microbiome, and the redundancy of multiple modules of a cohort falling within a cluster is potentially a further stabilizing force for the ecosystem. These findings agree with previous studies showing comparable communities and high functional redundancy across gut microbiome data sets^{55,67}.

Lastly, we identified species that acted as “hubs” or “bottlenecks” within the structure of cohort networks. These node types are important for maintaining network structure and thereby also potentially important species for community structure within the human gut microbiome³⁵. Notably, we found *Bacteroides* sp. were designated as both “hubs” and “bottlenecks” across all networks. Interestingly, *Bacteroides* sp. were also found to be the largest constituent of bacterial cliques and these cliques were the most retained across all cohorts. Additionally, of the 20 species from the 14 conserved associations found across networks, most were species belonging to *Bacteroides*. These findings suggest that *Bacteroides* sp. are important drivers of the ecosystem within the healthy human gut microbiome. Interestingly, previous studies have also designated *Bacteroides* sp., such as *Bacteroides fragilis* and *Bacteroides stercosis*, as potentially important (keystone) species within the human gut microbiome⁶⁸.

It is important to consider the limitations of our study. Our samples originated from different geographical locations and utilized different preparation procedures both of which are known to introduce biases^{24,69,70}. Another limitation is the presence of potential confounding variables within the cohorts such as age and sex. Additionally, due to the cross-sectional nature of our data we are only able to capture snapshots of the gut microbiome and are unable to examine the dynamics of the ecosystem. Furthermore, we utilized a reference-based mapping approach for taxonomic classification potentially causing our classifications to be limited by the genomes available. Finally, the constructed bacterial networks were undirected, and the study was non-mechanistic which prevents us from being able to examine the influence individual species have on one another (unidirectional ecological interactions).

In closing, we have demonstrated that bacterial communities across healthy human populations are similar in their organization and functional capacities. We have also revealed that positive associations regularly occur between taxonomically and functionally related species despite bacterial carriage differences, healthy human gut microbiomes across populations exhibit less variation (structural and functional) than previously believed. Our future research will build upon these findings to better understand how bacterial associations change within the disease microbiome. Also, by using the prevalent species, we can minimize the ‘noise’ of bacterial variation across hosts, especially since low prevalence species may ultimately be transient in nature⁴². This could be advantageous as it has been suggested that the most abundant organisms are the ones that act as “ecosystem engineers”⁵², and the study of these organisms would be important to understand how the microbiome responds to disturbances.

Materials and methods

Data acquisition. We utilized 606 WGS fecal samples (1.68 Tbp), which were obtained from four previously published human gut microbiome studies from four different healthy human populations (cohorts). Three cohort datasets were downloaded from the NCBI Sequence Read Archive (SRA): American¹⁵ (PRJNA48479; 202 samples), Indian⁷¹ (PRJNA397112; 106 samples), and European⁷² (PRJEB2054; 120 samples). The Japanese cohort dataset was downloaded from the DDBJ Sequence Read Archive (DRA): Japanese⁷³ (PRJDB4176; 178 samples) (Supplemental 13). All cohort sample groups had similar male-to-female frequencies, except for the European cohort (American: 0.53/0.47; Indian: 0.50/0.50; European: 0.34/0.66; Japanese: 0.56/0.44) (Supplemental 14).

Data pre-processing. Reads from all samples were first trimmed using Trimmomatic⁷⁴ (version 0.36) and then human reads were filtered using BowTie2⁷⁵ (version 5.4.0) and the GRCh38.p12 (https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.38/) human reference genome. After removal of human reads, 15.9 billion high-quality reads remained (Supplemental 15).

Read mapping and species-level taxonomic profiling. Reads were mapped to a collection of 10,839 bacterial reference strain genomes downloaded from RefSeq⁷⁶, using Bowtie2. The read mapping information was analyzed using a probabilistic framework based on a mixture model to estimate the relative copy number of each reference genome in a sample. This framework used an Expectation–Maximization (EM) algorithm to optimize the log-likelihood function associated with the model⁷⁷. The EM algorithm was found to be highly accurate when benchmarked using simulated WGS reads produced by WGSim (<https://github.com/lh3/wgsim>) (Supplemental 16). Sub-sampling and benchmark testing of sample read mapping counts showed that a read depth of 250,000 mapped reads at a noise threshold of $1e-5$ correlated well with samples mapping over 5 million mapped reads ($R^2 > 0.85$, Supplemental 17). Any bacterial strain found in a sample below $1e-5$ relative abundance was considered statistical noise and was dropped to an abundance of 0. Strains were then grouped by their species classification and their relative abundances were summed to produce species abundances.

Bacterial genome annotation and functional profiles. All bacterial reference genomes were functionally annotated in-house to create reference strain functional profiles. Before genome annotation, we utilized CheckM⁷⁸ (v1.0.13) to ensure that these reference genomes were mostly complete (Supplemental 18). Prodigal⁷⁹ (version 2.6.3) was used to identify genes, and generate protein sequence translations, which were then provided to InterProScan⁸⁰ (version 5.39-77.0) to find matches to protein families using the TIGRFAM⁸¹ (version 15.0) database. The functional profile for a bacterial strain was created by identifying the total number TIGRFAM matches to the strain, and subsequently converting these counts to relative abundances. The functional profile for a bacterial species was created separately for each cohort. This was computed by first finding the average genome abundance of each strain within the cohort, weighting the strain functional profiles based on these proportions, and then aggregating the resulting strain profiles. Each species functional profile was then CLR-transformed. CLR-transformation is defined as:

$$\text{clr}(x) = \left[\ln \frac{x_1}{g(x)}, \dots, \ln \frac{x_2}{g(x)}, \dots, \ln \frac{x_D}{g(x)} \right]$$

where x is the vector of species abundances within each sample, D is the total number of species. The geometric mean of vector x is defined as:

$$g(x) = \sqrt[D]{x_1 \times x_2 \times \dots \times x_D}$$

TIGRFAM functional annotations were obtained from TIGRFAMS_ROLE_LINK and TIGRFAM_ROLE_NAMES (ftp://ftp.jcvi.org/pub/data/TIGRFAMS/14.0_Release/).

Cohort sample functional profiling. A Simplified Annotation Format (SAF) file containing the bacterial chromosomal coordinates of TIGRFAMS (features) for all reference strains was provided to FeatureCounts⁸² (Subread package 2.0.0) to find the total features contained within sample reads. Counts of features were subsequently length normalized, summed, and re-normalized (by total) for each sample producing sample functional profiles. Protein families were grouped by their TIGRFAM role, and their relative abundances were aggregated and CLR-transformed to generate the cohort functional role profiles (CFRP). Roles that were a different sign (+/–) in one cohort, when compared to all other cohorts, were considered different (elevated/reduced).

Construction of bacterial association networks. For each cohort, a sample-taxa matrix was constructed containing the relative abundances of the signature species in each sample. The bacterial association network for a cohort was constructed from its CLR transformed sample-taxa matrix using the GGM framework. In each case, a sparse precision matrix was computed using the R³³ huge⁸⁴ package, and this matrix formed the adjacency matrix of the association network. The tuning parameter ρ in the l1-penalty model for sparse precision matrix estimation was chosen using the stability approach to regularization (StARS) method⁸⁵. In order to reduce the number of false positives, the estimated sparse precision matrix Ω was processed further using a bootstrap method as follows: r bootstrap datasets, each with n samples, were generated from the original CLR-transformed matrix by random sampling with replacement. A sparse precision matrix was estimated from each bootstrap dataset using the same previously chosen value of the tuning parameter ρ used to estimate Ω .

The final precision matrix Ω' is derived from Ω as follows: (a) if $\Omega[i,j]=0$, then $\Omega'[i,j]=0$. (b) if $\Omega[i,j] \neq 0$, then $\Omega'[i,j]=\Omega[i,j]$ if the entry $[i,j]$ is non-zero in at least $f \times r$ precision matrices estimated from the bootstrap datasets. Otherwise $\Omega'[i,j]=0$. Thus, Ω' is at least as sparse as Ω . Partial Correlation matrix, P , was calculated as:

$$P_{[i,j]} = \frac{-\Omega'_{[i,j]}}{\sqrt{\Omega'_{[i,i]} \times \Omega'_{[j,j]}}}$$

The value f is a preset threshold ($0 \leq f \leq 1$). We used $r=50$ (bootstrap replicates) and $f=0.8$ (e.g. association must be non-zero $\geq 80\%$ of the time) in our analysis. Partial correlation matrices were parsed using python and all associations below a magnitude of 0.01 were considered statistical noise and removed.

Network property, clique, and module analysis. For each cohort network, the following properties were computed using NetworkX⁸⁶ (version 2.4): average shortest path length (ASPL), transitivity, modularity, degree assortativity, degree centrality, betweenness centrality, and genera assortativity. The ASPL (α) is defined as:

$$\alpha = \frac{\sum_{s,t \in V} D[s,t]}{n(n-1)}$$

where V is the set of nodes in the graph (G), $D[s,t]$ is the shortest path from s to t , and n is the total number of nodes in G . The transitivity (T) of a network is the fraction of all possible triangles present in the graph, and is defined as:

$$T = 3 \frac{\#triangles}{\#triads}$$

triangles are a clique (a subset of nodes within a network where each node is adjacent to all other nodes within the subset) of three nodes, and triads are the count of connected triples (three nodes xyz with edges (x,y) and (y,z) where the edge (x,z) can be present or absent)^{86,87}. Modularity (Q) is defined as:

$$Q = \frac{1}{2m} \sum_{i,j} \left(A[i,j] - \frac{k_i k_j}{2m} \right) \delta(C_i, C_j)$$

where A is the adjacency matrix of graph (G), m is the total number of edges, k_i is the degree of node i , and $\delta(C_i, C_j)$ is 1 if i and j (node pair) are in the same community or 0 if in different communities^{87,88}. Assortative mixing is a predilection of nodes to form connections with other nodes that are like (assortative) or unlike (disassortative) themselves. We measured node mixing preference according to node degree (degree assortativity) and node genus classification (genera assortativity). Degree assortativity is calculated using the standard Pearson correlation coefficient:

$$r = \frac{\sum_{xy} xy(D[x,y] - a_x b_y)}{\sigma_a \sigma_b}$$

where D is the joint probability distribution matrix, $D[x,y]$ is the fraction of all edges in the graph that connects nodes with degree values x and y , a_x and b_y are the fraction of edges that start and end at nodes with values x and y , and σ_a and σ_b are the standard deviations of the distributions a_x and b_y . The value of r can be any value between -1 (perfect disassortativity) and 1 (perfect assortativity). Genera assortativity is defined as:

$$r = \frac{\text{Tr}Q - Q^2}{1 - Q^2}$$

where Q is the joint probability distribution matrix whose elements are $Q[i,j]$ (the fraction of all edges in the graph that connects nodes of genus type i to genus type j), Tr is the trace of the matrix Q , and $\|Q\|$ signifies the sum of all elements of the matrix Q ⁸⁹.

Modules within each network were found utilizing the *label_propagation_communities* algorithm, based on the asynchronous label propagation algorithm (aLPA)⁴⁵ from NetworkX. To quantify the ability of the aLPA to partition the data, we utilized the *performance* function NetworkX. Performance (p) is defined as:

$$p = \frac{a+b}{c}$$

where a is the total intra-module edges, b is the total inter-module non-edges, and c is the total potential edges⁹⁰. Monte Carlo simulations were utilized to test for statistical significance of network property differences (see Supplemental). Three member cliques and modules within each network were found using NetworkX. Module functional profiles (MFP) were created by aggregating the functional profiles of species contained within each module.

Network node centrality (“hubs” and “bottlenecks”) analysis. Degree centrality is defined as the degree (total edges) of a node. The node within each network module exhibiting the highest degree central-

ity was designated as a module “hub”. If two or more species were found to have equal degree centrality then centrality measurements of those nodes were re-computed in context of the entire network. The top ten nodes exhibiting the highest betweenness centrality within each network were designated as “bottlenecks”. To find “bottleneck” species, betweenness centrality was computed for each node. Betweenness centrality is defined as:

$$C_B(u) = \sum_{s,t \in V} \frac{\sigma(s,t|u)}{\sigma(s,t)}$$

where the betweenness centrality of a node (v) is the sum of the fraction of all-pairs shortest paths that pass through v , V is the set of all nodes, $\sigma(s,t)$ is the number of shortest paths (s,t)-paths, and $\sigma(s,t|v)$ is the number of those paths passing through node v other than s,t ⁹¹.

Data and script availability

All scripts and data from this study is available for download at github: (<https://github.com/syooseph/YoosepHLab/tree/master/MicrobiomeNetworks/HealthyPopulations>).

Received: 27 September 2020; Accepted: 20 January 2021

Published online: 02 February 2021

References

1. Sender, R., Fuchs, S. & Milo, R. Revised estimates for the number of human and bacteria cells in the body. *PLOS Biol.* **14**, e1002533 (2016).
2. Kho, Z. Y. & Lal, S. K. The human gut microbiome—A potential controller of wellness and disease. *Front. Microbiol.* **9**, 1835 (2018).
3. Kostic, A. D., Xavier, R. J. & Gevers, D. The microbiome in inflammatory bowel disease: Current status and the future ahead. *Gastroenterology* **146**, 1489–1499 (2014).
4. Thaiss, C. A., Zmora, N., Levy, M. & Elinav, E. The microbiome and innate immunity. *Nature* **535**, 65–74 (2016).
5. Das, B. & Nair, G. B. Homeostasis and dysbiosis of the gut microbiome in health and disease. *J. Biosci.* **44**, 117 (2019).
6. Shreiner, A. B., Kao, J. Y. & Young, V. B. The gut microbiome in health and in disease. *Curr. Opin. Gastroenterol.* **31**, 69–75 (2015).
7. Petersen, C. & Round, J. L. Defining dysbiosis and its influence on host immunity and disease: How changes in microbiota structure influence health. *Cell. Microbiol.* **16**, 1024–1033 (2014).
8. Karlsson, F. H. *et al.* Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* **498**, 99–103 (2013).
9. Koren, O. *et al.* Human oral, gut, and plaque microbiota in patients with atherosclerosis. *Proc. Natl. Acad. Sci.* **108**, 4592–4598 (2011).
10. Karlsson, F. H. *et al.* Symptomatic atherosclerosis is associated with an altered gut metagenome. *Nat. Commun.* **3**, 1245 (2012).
11. Chatelier, L. M. *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**, 541–546 (2013).
12. Franzosa, E. A. *et al.* Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat. Microbiol.* **4**, 293–305 (2019).
13. Becker, C., Neurath, M. F. & Wirtz, S. The intestinal microbiota in inflammatory bowel disease. *ILAR J.* **56**, 192–204 (2015).
14. Kostic, A. D. *et al.* *Fusobacterium nucleatum* potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. *Cell Host Microbe* **14**, 207–215 (2013).
15. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
16. Johnson, A. J. *et al.* Daily sampling reveals personalized diet–microbiome associations in humans. *Cell Host Microbe* **25**, 789–802. e5 (2019).
17. Villmones, H. C. *et al.* Species level description of the human ileal bacterial microbiota. *Sci. Rep.* **8**, 1–9 (2018).
18. Gevers, D. *et al.* The treatment-naïve microbiome in new-onset Crohn’s disease. *Cell Host Microbe* **15**, 382–392 (2014).
19. David, L. A. *et al.* Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**, 559–563 (2014).
20. Zhou, J., Deng, Y., Luo, F., He, Z. & Yang, Y. Phylogenetic molecular ecological network of soil microbial communities in response to elevated CO₂. *MBio* **2**, e00122–e211 (2011).
21. Lupatini, M. *et al.* Network topology reveals high connectance levels and few key microbial genera within soils. *Front. Environ. Sci.* **2**, 10 (2014).
22. Eiler, A., Heinrich, F. & Bertilsson, S. Coherent dynamics and association networks among lake bacterioplankton taxa. *ISME J.* **6**, 330–342 (2012).
23. Kara, E. L., Hanson, P. C., Hu, Y. H., Winslow, L. & McMahon, K. D. A decade of seasonal dynamics and co-occurrences within freshwater bacterioplankton communities from eutrophic Lake Mendota, WI, USA. *ISME J.* **7**, 680–684 (2013).
24. Shetty, S. A., Hugenholtz, F., Lahti, L., Smidt, H. & de Vos, W. M. Intestinal microbiome landscaping: Insight in community assemblage and implications for microbial modulation strategies. *FEMS Microbiol. Rev.* **41**, 182–199 (2017).
25. Gould, A. L. *et al.* Microbiome interactions shape host fitness. *Proc. Natl. Acad. Sci.* **115**, E11951–E11960 (2018).
26. Hibbing, M. E., Fuqua, C., Parsek, M. R. & Peterson, S. B. Bacterial competition: Surviving and thriving in the microbial jungle. *Nat. Rev. Microbiol.* **8**, 15–25 (2010).
27. Fox, G. E., Magrum, L. J., Balcht, W. E., Wolfef, R. S. & Woese, C. R. Classification of methanogenic bacteria by 16S ribosomal RNA characterization (comparative oligonucleotide cataloging/phylogeny/molecular evolution). *Evolution (N.Y.)* **74**, 4537–4541 (1977).
28. Venter, J. C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66 (2004).
29. Větrovský, T. & Baldrian, P. The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PLoS ONE* **8**, e57923 (2013).
30. Edgar, R. C. Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences. *PeerJ* **6**, 1–29 (2018).
31. Ranjan, R., Rani, A., Metwally, A., McGee, H. S. & Perkins, D. L. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochem. Biophys. Res. Commun.* **469**, 967–977 (2016).
32. Laudadio, I. *et al.* Quantitative assessment of shotgun metagenomics and 16S rDNA amplicon sequencing in the study of human gut microbiome. *Omi. A J. Integr. Biol.* **22**, 248–254 (2018).
33. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome datasets are compositional: And this is not optional. *Front. Microbiol.* **8**, 1–6 (2017).
34. Friedman, J. & Alm, E. J. Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* **8**, e1002687 (2012).
35. Layeghifard, M., Hwang, D. M. & Guttman, D. S. Disentangling interactions in the microbiome: A network perspective. *Trends Microbiol.* **25**, 217–228 (2017).

36. Aitchison, J. The statistical analysis of compositional data. *J. R. Stat. Soc. Ser. B* **44**, 40 (1982).
37. Kurtz, Z. D. *et al.* Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.* **11**, 1–25 (2015).
38. Friedman, J., Hastie, T. & Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–441 (2008).
39. Falony, G. *et al.* Population-level analysis of gut microbiome variation. *Science* **352**, 560–564 (2016).
40. Efron, B. & Tibshirani, R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat. Sci.* **1**, 54–75 (1986).
41. Su, W., Bogdan, M., Candès, E. & Candès, E. False discoveries occur early on the lasso path. *Ann. Stat.* **45**, 2133–2150 (2017).
42. Saunders, A. M., Albertsen, M., Vollertsen, J. & Nielsen, P. H. The activated sludge ecosystem contains a core community of abundant organisms. *ISME J.* **10**, 11–20 (2016).
43. Tsvetovat, M. & Kouznetsov, A. Social network analysis for startups. *Zhurnal Eksperimental'noi i Teoreticheskoi Fiziki* (2011).
44. Stadtfeld, C., Takács, K. & Vörös, A. The emergence and stability of groups in social networks. *Soc. Netw.* **60**, 129–145 (2020).
45. Cordasco, G. & Gargano, L. Community detection via semi-synchronous label propagation algorithms. *2010 IEEE Int. Work. Bus. Appl. Soc. Netw. Anal. BASNA 2010* (2010). <https://doi.org/10.1109/BASNA.2010.5730298>.
46. Prettjohn, B. J., Berryman, M. J. & McDonnell, M. D. Methods for generating complex networks with selected structural properties for simulations: A review and tutorial for neuroscientists. *Front. Comput. Neurosci.* **5**, 11 (2011).
47. Guimerà, R., Sales-Pardo, M. & Amaral, L. A. N. Modularity from fluctuations in random graphs and complex networks. *Phys. Rev. E* **70**, 25101 (2004).
48. Trosvik, P. & de Muinck, E. J. Ecology of bacteria in the human gastrointestinal tract—Identification of keystone and foundation taxa. *Microbiome* **3**, 44 (2015).
49. Verster, A. J. & Borenstein, E. Competitive lottery-based assembly of selected clades in the human gut microbiome. *Microbiome* **6**, 186 (2018).
50. Berry, D. & Widder, S. Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Front. Microbiol.* **5**, 219 (2014).
51. Foster, K. R. & Bell, T. Competition, not cooperation, dominates interactions among culturable microbial species. *Curr. Biol.* **22**, 1845–1850 (2012).
52. Nemergut, D. R. *et al.* Patterns and processes of microbial community assembly. *Microbiol. Mol. Biol. Rev.* **77**, 342–356 (2013).
53. Hamilton, W. D. The genetical evolution of social behaviour. I. *J. Theor. Biol.* **7**, 1–16 (1964).
54. Hardin, G. The competitive exclusion principle. *Science* **131**, 1292–1297 (1960).
55. Jackson, M. A. *et al.* Detection of stable community structures within gut microbiota co-occurrence networks from different human populations. *PeerJ* **6**, e4303 (2018).
56. Darcy, J. L. *et al.* A phylogenetic model for the recruitment of species into microbial communities and application to studies of the human microbiome. *ISME J.* **14**, 1359–1368 (2020).
57. Pacheco, A. R., Moel, M. & Segrè, D. Costless metabolic secretions as drivers of interspecies interactions in microbial ecosystems. *Nat. Commun.* **10**, 103 (2019).
58. Chase, J. M. & Leibold, M. A. Spatial scale dictates the productivity–biodiversity relationship. *Nature* **416**, 427–430 (2002).
59. Zarrinpar, A., Chaix, A., Yooseph, S. & Panda, S. Diet and feeding pattern affect the diurnal dynamics of the gut microbiome. *Cell Metab.* **20**, 1006–1017 (2014).
60. Mark Welch, J. L., Hasegawa, Y., McNulty, N. P., Gordon, J. I. & Borisy, G. G. Spatial organization of a model 15-member human gut microbiota established in gnotobiotic mice. *Proc. Natl. Acad. Sci.* **114**, E9105–E9114 (2017).
61. Fung, T. C., Artis, D. & Sonnenberg, G. F. Anatomical localization of commensal bacteria in immune cell homeostasis and disease. *Immunol. Rev.* **260**, 35–49 (2014).
62. Donaldson, G. P., Lee, S. M. & Mazmanian, S. K. Gut biogeography of the bacterial microbiota. *Nat. Rev. Microbiol.* **14**, 20–32 (2016).
63. Stachowicz, J. J. Mutualism, facilitation, and the structure of ecological communities. *Bioscience* **51**, 235 (2001).
64. Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K. & Knight, R. Diversity, stability and resilience of the human gut microbiota. *Nature* **489**, 220–230 (2012).
65. Lynd, L. R., Weimer, P. J., van Zyl, W. H. & Pretorius, I. S. Microbial cellulose utilization: Fundamentals and biotechnology. *Microbiol. Mol. Biol. Rev.* **66**, 72 (2002).
66. Turroni, F. *et al.* Glycan cross-feeding activities between bifidobacteria under in vitro conditions. *Front. Microbiol.* **6**, 1030 (2015).
67. Hall, C. V. *et al.* Co-existence of network architectures supporting the human gut microbiome. *iScience* **22**, 380–391 (2019).
68. Fisher, C. K. & Mehta, P. Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. *PLoS ONE* **9**, e102451 (2014).
69. Jones, M. B. *et al.* Library preparation methodology can influence genomic and functional predictions in human microbiome research. *Proc. Natl. Acad. Sci.* **112**, 14024–14029 (2015).
70. Lahti, L., Salojärvi, J., Salonen, A., Scheffer, M. & de Vos, W. M. Tipping elements in the human intestinal ecosystem. *Nat. Commun.* **5**, 4344 (2014).
71. Dhakan, D. B. *et al.* The unique composition of Indian gut microbiome, gene catalogue, and associated fecal metabolome deciphered using multi-omics approaches. *Gigascience* **8**, giz004 (2019).
72. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).
73. Yachida, S. *et al.* Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nat. Med.* **25**, 968–976 (2019).
74. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
75. Langmead, B. & Salzberg, S. Bowtie2. *Nat. Methods* **9**, 357–359 (2013).
76. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
77. Xia, L. C., Cram, J. A., Chen, T., Fuhrman, J. A. & Sun, F. Accurate genome relative abundance estimation based on shotgun metagenomic reads. *PLoS ONE* **6**, e27992 (2011).
78. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
79. Hyatt, D. *et al.* Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* **11**, 119 (2010).
80. Jones, P. *et al.* InterProScan 5: Genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
81. Haft, D. H. TIGRFAMs: A protein family resource for the functional identification of proteins. *Nucleic Acids Res.* **29**, 41–43 (2001).
82. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
83. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, 2017).
84. Zhao, T., Liu, H., Roeder, K., Lafferty, J. & Wasserman, L. The huge package for high-dimensional undirected graph estimation in R. *J. Mach. Learn. Res.* **13**, 6 (2016).

85. Liu, H., Roeder, K. & Wasserman, L. Stability approach to regularization selection (stars) for high dimensional graphical models. *Advances in Neural Information Processing Systems* (2010).
86. Hagberg, A., Swart, P. & Chult, D. S. *Exploring network structure, dynamics, and function using NetworkX*. No. LA-UR-08-05495; LA-UR-08-5495 (Los Alamos National Lab. (LANL), Los Alamos, 2008).
87. Newman, M. E. J. *Networks: An Introduction* 168–234 (Oxford University Press, Oxford, 2010).
88. Newman, M. E. J. Modularity and community structure in networks. *Proc. Natl. Acad. Sci.* **103**, 8577–8582 (2006).
89. Newman, M. E. J. Mixing patterns in networks. *Phys. Rev. E* **67**, 26126 (2003).
90. Fortunato, S. Community detection in graphs. *Phys. Rep.* **486**, 75–174 (2010).
91. Brandes, U. A faster algorithm for betweenness centrality*. *J. Math. Sociol.* **25**, 163–177 (2001).

Author contributions

Study design: M.L., S.A.-D.H., S.Y.; data analysis: M.L., S.A.-D.H., S.Y.; manuscript writing: M.L., S.A.-D.H., S.Y.

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/s41598-021-82449-0>.

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021