# ARRiVAL: Analyzing Ribosomal RNA in Variable microbiome Architecture Layouts

Finding disease-specific shifts in the human gut microbiome through correlation network-based modeling

Caroline Ting, Jana Reiser, Kailey Hua, Maxine Lui

Carnegie Mellon University, Pittsburgh, PA 15213

April 29, 2024

## Abstract

The gut microbiome has been shown to have great influence over the human digestive system and the immune system, as well as neuronal well-being, metabolism, and overall health. Studies have shown that the gut microbiome is capable of positively influencing human health, from protecting against pathogenic invasion to strengthening the immune system. Microbiota encode more genes than their human hosts, and thus control a variety of metabolic functions beyond the scope of our own systems. An alteration of the diversity and abundance of microorganisms in the gut microbiome has been implicated in a variety of diseases, most commonly irritable bowel syndrome (IBS) and inflammatory bowel disease (IBD), where IBD is a precursor to colorectal cancer (CRC). In this paper, we present a topological network analysis of the microbiome of patients suffering from IBS, IBD, and CRC. All of which are diseases that are increasing significantly in recent years and are associated with inflammation in the gastrointestinal tract (GIT). We analyze the biodiversity and interconnectedness of the microbiome of patients with IBS, IBD, and CRC to gain a better understanding of the resilience and stability of the microbiomes and the factors that lead to inflammation of the GIT. Possible applications of these analyses include microbiota-targeted therapeutic treatments for overgrown or missing bacteria, as well as tracking the severity of the disease progression. We also found the most significant bacteria present in patients with IBS, IBD, and CRC, which act as biomarkers for the diseases for non-invasive diagnosis.

## 1. Introduction

The *gut microbiome* is defined as the class of microorganisms inhabiting the human digestive tract, and its imbalance has been linked with many digestive conditions. Extensive research has been conducted to study the relationship between the gut microbiome and gastrointestinal conditions. There is also ongoing research for gut microbiome-targeted therapeutics for treating many of these conditions, and to stabilize it to a healthy balance of microorganisms (1).

In current literature, studies have used network-based approaches to characterize the abundance and biodiversity of microbiota in different groups of patients. The ability to understand the differences between the gut microbiome of healthy individuals and diseased states will allow us to design more accurate microbiota-targeted therapeutics as well as gain a larger understanding of the affects of gastrointestinal conditions (5).

We focus on a network analysis of the connections between taxa in the gut microbiome of IBS and IBD, two common gastrointestinal conditions, and CRC, a cancer which is intimately associated with the gut microbiome, as it makes up much of the tumor's microenvironment (9). Our work shows the shifts in abundance and interconnectedness of the gut microbiome in these three disease classes, and can aid researchers in microbiota-based disease class annotation.

### 1.1. Gut microbiome and the specific disease classes

Patients with IBS, IBD, and CRC seem to suffer *dysbiosis*, an imbalance of gut flora that happens due to the overgrowth of particular conspecifics, individuals from the same taxonomic family. This may reduce microbial diversity and trigger inflammatory changes.

#### 1.1.1. Irritable bowel syndrome

Irritable bowel syndrome (IBS) is a functional gastrointestinal disorder (FGID) with symptoms of abdominal pain and discomfort and is associated with psychiatric disorders such as anxiety and depression. IBS is a dysregulation of communication between the gut-brain axis and a change in gut microbiome composition (7).

In particular, IBS occurs commonly alongside other gastrointestinal (GI) conditions; many patients are also diagnosed with gastroesophageal reflux disorder (GERD). Furthermore, research on the gut-brain axis indicate that microbiota plays a key role in hormonal management, and in particular, stress response is linked to pathogenic bacteria *Pseudomonas aeruginosa* and *Campylobacter jejuni*. Laboratory research has found key differences in the gut microbiota of IBS patients and healthy individuals, particularly upregulated *Ruminococcus gnavus* and *Lachnospiraceae* and downregulated *Barnesiella intestinihominis* and *Coprococcus catus*. However, studies have thus far been ineffective in finding notable differences in microbiota levels.

### 1.1.2. Inflammatory bowel disease (IBD)

Inflammatory bowel disease (IBD) is believed to occur due to a combination of genetic risk factors and environmental factors. Prior research indicates a bidirectional relationship between the state of the gut microbiome and the progression of the disease. IBD is an umbrella term for chronic disorders and usually includes Crohn's Disease and ulcerative colitis, which is usually restricted to the colon and may be related to colorectal cancer (9).

Research has been done with low-resolution 16S rRNA-seq data, but there is a lack of current literature using high-resolution data, and it is unclear if the gut microbiome is the cause or symptom of IBD . The Integrative Human Microbiome Project (IHMP) studied individual IBD patients and found that certain microbial populations were downregulated in patients suffering from the disease. There were increases in the abundance of *Cloacibacterium* and *Tissierellaceae* alongside a decrease in *Neisseria* (2). Other bacteria kingdoms, such as Fungi, have been associated with IBD and intestinal permeability is highly associated with IBD, and may be affected by the characteristics of the gut microbiome.

### 1.1.3. Colorectal Cancer (CRC)

Colorectal cancer (CRC) is an unique cancer as studies have shown that there are changes in the microbiota as the disease progresses. There is ongoing research seeking to target the gut microbiome in order to treat CRC through microbiota implantations, or to use it as a non-invasive biomarker for CRC screening in high-risk patients (9).

Large-scale multicohort analysis of the CRC microbiome has yielded the finding of a core microbiome, a set of flora with tumorgenesis-related functions that may contribute to dysbiosis. In particular, researchers identified *Fusobacterium nucleatum* and *Peptostreptococcus anaerobius* to be bacteria highly involved in tumor initiation. These species are enriched in CRC while bacteria associated with probiotics are depleted. Furthermore, bacteria have the ability to colonize tumors, enriching the CRC microbiome. However, the variability of the microbiome across patients has made it difficult to effectively identify potential therapeutic targets.

## 1.2. Topological network analysis of microbiome-based graphs

Analysis of networks based on the microbiome of patients in certain disease cohorts can provide key insights on the differences between microbiota as well as the abundance and correlation between taxa. We first define a *microbiota network* to be a graph where nodes correspond to taxa on a hierarchical level in the phylogenetic tree, for example, elements of the family level. Edges are undirected and exist between nodes if the two taxa co-occur at a significant level. This graph is not necessarily corrected nor acyclic, but has no multi-edges and no self-edges. Formally, we have a microbiota network $G = (V, E)$ where $V$ is some phylogenetic level and $E$ is tuples of distinct elements of $V$. In particular, we formalize the the notion of difference between diseases by analyzing the following attributes of a microbiota network:

1. **Interconnectedness**, i.e. a measure of how many bacteria tend to be upregulated together and whose expression may be correlated,
2. **Connected motifs**, subgraphs such as a complete graph on 4 vertices ($K_4$) that occur more often than random,
3. **Clustering degree**, or a representation of the probability that two neighbors of a node are connected, or how interconnected the graphs are,
4. **The cumulative distribution function**, or a measure of how many neighbors a given conspecies may have, and
5. **Alpha biodiversity**, the species richness on a local scale, and
6. **Edge-weighting**, or average shortest path analysis to identify key species.

We compute many of these metrics for multiple taxonomic hierarchical ranks to gain a richer understanding of the differences between each cohort. These attributes give us a in-depth quantitative analysis of the interconnectedness and relevance of each conspecies in the microbiome of each cohort. Furthermore, between the cohorts, we analyze:

1. **Beta diversity**, a measure of how different two regions are, and
2. **Core microbiota**, an intersection of all microbiota graphs.

## 2. Methods

The Python implementation of the network analysis and the Docker-dependent bash script for BLAST search against the Ribosomal Databse Project can be found on the GitHub repository linked under **Code availability**. All computation is done using $\ell_1$-normalized data to account for biases, differences in number of patients and read count per patient, and potential batch effect. The $\ell_1$ normalization of data is done with respect to its norm, i.e. for a vector $\vec{v}$, define $\ell_1$-normalized $\vec{v}' = \frac{\vec{v}}{||\vec{v}||}$.

### 2.1. Taxonomic identification

The data we used to study the differences between the microbiota in *16S ribosomal RNA*, abbreviated 16S rRNA. It represents the 30S subunit of a prokaryotic ribosome (small subunit ribosomal ribonucleic acid) which evolves slowly, and is therefore imperative for reconstructing phylogenetics (3). It is the most common housekeeping genetic marker and is used in state-of-the-art bacterial identification. The IBS, IBD, and CRC data is longitudinal multi-omics data accounting for both the gut microbiome and the transcriptome. This resulted in the following data statistics:

| | IBS | IBD | CRC |
|---|---|---|---|
| Number of patients | 87 | 177 | 42 |
| Number of reads | 229016 | 31928 | 84924 |
| Reads per patient | 2632 | 180 | 2022 |

Figure 1. General dataset statistics prior to RDP Classifier BLAST search.

We note that IBD is characterized by a relatively low reads per patient count, implying that there may be less raw microbiota in IBD pre-normailzation. This is supported by previous studies noting the loss in both diversity and raw number of bacteria in the IBD microbiome.

In order to derive taxonomic, or phylogenetic, information from 16S rRNA-sequencing reads (16S rRNA-seq reads), we used a Basic Local Alignment Search Tool (BLAST) against the Ribosomal Database Project (RDP) with confidence level 0.5, as the rRNA-seq reads were relatively short (on average 100 nucleotides). This was done on a total of 306 patients spanning all 3 disease cohorts using a local installation of `RDPTools`, a repository of command-line based tools including the RDP Classifier. This software tool takes rRNA-seq from bacterial and archaeal genes, and fungal LSU and ITS sequences, and matches them against their correct taxonomy models. The RDP Classifier is a *naive Bayesian classifier*, a supervised learning algorithm which is commonly used in classification tasks (8). It applies Bayes' rule, i.e. for two events $A, B$, $\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}$. The RDP Classifier thus assumes independence between features on the phylogenetic tree and calculates the probability that the rRNA-seq read belongs to a given class. We eliminate sequences with ambiguous bases and set the minimum accepted overlap to 10 bases (`-N -o 10`).

After conducting BLAST search using the 16S rRNA-seq reads, we filtered our data for the top 40% most confident hits. From here, we constructed a dictionary of the conspecifics at each hierarchical level of Linnaean taxonomy, which is:

1. **Domain**, the highest classification rank–the "three domains of life," Archaea, Bacteria, and Eucarya,
2. **Phylum**, a class of organisms with high morphological similarity,

3. **Class**, a group of organisms that share some common attribute,
4. **Order**, which is defined to be a group of closely related families,
5. **Family**, a group of genera that usually share a common attribute, and
6. **Genus**, a group of species marked by a common characteristic, whose relationships tend to be obscure due to an incomplete understanding of their structure or development.

After performing BLAST search against each cohort, we found the following properties for the union of gut microbiota:

| | IBS | IBD | CRC |
|---|---|---|---|
| Domain | 3 | 3 | 3 |
| Phylum | 56 | 45 | 59 |
| Class | 123 | 100 | 130 |
| Order | 258 | 183 | 274 |
| Family | 581 | 351 | 614 |
| Genus | 2313 | 1090 | 2501 |

Figure 2. Representation of high confidence level taxons in each hierarchical level in IBS, IBD, and CRC after BLAST search of 16S rRNA-seq reads.

As our data is normalized in all analyses, we account for differences in total read count. We chose to focus our analysis on the family level of taxa as it provided the most nuanced microbiota network. The network constructed for taxa on order and above were too similar, while those constructed for the genus level were too fine grained and noisy.

### 2.2. Biodiversity analysis

*Beta diversity* is defined to be the ratio of regional (or overall) diversity to local species diversity in a confined ecosystem, or in this case, a microbiome. In essence, it quantifies the number of different species between two regions. There are two models for computing beta diversity defined as the Jaccard Index and Bray–Curtis dissimilarity.

The *Jaccard Index* produces a value between 0 and 1 indicating the percent similarity between two populations. For a set of all species $\Omega$ and populations $A, B \subseteq \Omega$, the Jaccard Index is a function $J : \Omega \times \Omega \to [0, 1]$ defined via

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}.$$

The weighted *Jaccard distance* $J_D : \Omega \times \Omega \to [0, 1]$ is then defined to be $1 - J(A, B)$, and is more intuitively computed as follows:

$$J_D(A, B) = 1 - \frac{\sum_{s \in A \cap B} \min(A[s], B[s])}{\sum_{s \in A \cup B} \max(A[s], B[s])}.$$

The *Bray-Curtis dissimilarity* quantifies the dissimilarity between two sites based on the count of species on each site. It is called a dissimilarity rather than a distance as it does not satisfy the triangle inequality, a key property of all distance functions. We represent this as a function $BC : \Omega \times \Omega \to [0, 1]$ where

$$BC(A, B) = 1 - \frac{2 \sum_{s \in A \cap B} \min(A[s], B[s])}{\sum_{s \in A \cup B} A[s] + B[s]}.$$

In both of these metrics, lower values close to 0 means that the sites $A, B$ are highly similar, while values close to 1 indicate dissimilarity.

The alpha diversity, a measure of how diverse a local region is, is computed using Simpson's Index of Diversity $S : \Omega \to [0, 1]$ using the formula

$$S(A) = 1 - \frac{\sum_{s \in A} A[s] \cdot (A[s] - 1)}{n \cdot (n - 1)}$$

Where $n$ is the total number of species in $A$. $S(A)$ then ranges from 0 to 1, with values closer to 0 indicating lower diversity and values at 1 indicating infinite diversity. In this case, the index value represents the probability that two randomly selected individuals from a sample belong to *different* species. In some existing literature, this may be referred to as the *complement* of Simpson's index.

### 2.3. Network modeling

To construct the microbiota networks, we make a graph $G = (V, E)$ with $V$ representing a phylogenetic level and $E = \{\{u, v\} : u \neq v \in V, d(u, v) > 2a\}$, where $a$ is the average co-occurrence of any two vertices. In other words, we construct an edge between two nodes in the graph if they co-occur at twice the average co-occurrence in all patients in the network. The function $d : V \times V \to \mathbb{N}$ measures how many patients the taxa co-occur in, and $a$ is computed as the average of all $d$, or

$$a = \frac{2 \sum_{u \neq v \in V} d(u, v)}{n(n - 1)}$$

where $n = |V|$ is the number of nodes. This helps overcome the batch effects of different datasets by using a twice-the-average heuristic rather than a constant cutoff, as some samples have on average more total microbiota.

### 2.3.1. Edge-weighting heuristic

For visualization purposes, edges $e = \{u, v\} \in E$ in a graph defined as $G = (V, E)$ were weighted such that more correlated taxa had a darker edges. More specifically, we computed the number of patients $p$ that $u, v$ were upregulated in, as well as the minimum number of patients $p_0$. This gave us that the weight of $e$ is $(10(p - p_0))^2$.

In the graph analysis, we assigned edges weights $0 \leq w \leq 1$, where a lower weight corresponds to more correlation, and thus a *shorter path.* This was computed using the formula $w = e^{-p}$, which is trivially bounded by the interval $[0, 1]$, and inversely correlated with $p$.

To compute the characteristic path length of a network, or the average shortest path length, we find

$$d = \sum_{s \neq t \in V} \frac{d(s,t)}{n(n-1)}$$

where $n = |V|$ and $d(s, t)$ denotes the distance from the shortest path from taxa $s$ to $t$. Dividing by $n(n - 1)$ effectively normalizes this value. This was computed on a graph with all isolated nodes removed, as there are no paths to and from those nodes. The weighted characteristic path length was calculated where $d(s, t)$ is the sum of weights in the shortest path. Note that the unweighted characteristic path length is just computed for edge-weight = 1 for all edges.

In measuring the average weighted degree, we returned to a weighting approach that assigned higher values to more correlated taxa, i.e. the percent of files they co-occurred in, or $w = \frac{p}{\text{num. patients}}$. Then a higher average weighted degree indicates more correlation, and this was computed using the formula

$$a = \frac{1}{n} \sum_{v \in V} \sum_{u \in N(v)} w_{u,v}.$$

## 2.4. Network topology

The *cumulative distribution function* (CDF) of a random variable $X$ evaluated at some value $x$ is the probability that $X$ will take on a value $\leq x$. For some CDF defined with respect to $X$, $F_X$, we define

$$F_X(x) = \Pr(X \leq x)$$

and on a semi-closed interval $a, b$,

$$\Pr(a < X \leq b) = F_X(b) - F_X(a).$$

Markov's inequality states that for $x \neq 0$,

$$\Pr(X \geq x) \leq 1 - \frac{E(X)}{x},$$

which implies that the CDF has the property

$$F_X(x) \approx 1 - \Pr(X \geq x) \geq \frac{E(X)}{x}.$$

Earlier, we presented the CDF where the random variable $X$ indicates the degree of a given node in the microbiota graph. However, since $X$ is not continuous (it can only take on discrete values in $\mathbb{N}$), we are unable to determine the probability density of the degrees of the nodes in the microbiota graph by taking the derivative.

Taxa with higher degree indicate that it may have correlation with a larger variety of taxa, or interact more with other taxa. These upregulated taxa of high-degree are likely biomarkers for disease cohorts.

The stability score was computed using the abundance-weighted mean interaction strength ($wMIS_i$ index). For a vertex (or taxa) $i$, we compute

$$wMIS_i = \frac{\sum j \neq i b_j \cdot |R_{ij}|}{\sum_{j \neq i} b_j},$$

where $b_j$ is the relative abundance of some taxa $j$ (i.e. number of species of $j$ divided by the total number of species) and $R_{ij}$ is the Spearmann correlation coefficient. We compute $R_{ij}$ as

$$R_{ij} = \frac{\text{cov}(i, j)}{\sigma_i \cdot \sigma_j}$$

where $\text{cov}(i, j)$ is covariance and $\sigma$ is standard deviation. Using the $wMIS_i$ index for each conspecies $i$, we can then compute the stability score of a network $G = (V, E)$ with core nodes $C \subseteq V$ via the formula

$$S(G) = \frac{\sum_{i \in C} wMIS_i}{\sum i \in V wMIS_i}.$$

We chose the core nodes to be the nodes who had twice the average number of neighbors, or the taxa that co-occurred with other taxa more frequently, and therefore were the key defining microbiota of the network.

### 2.5. Motif search and randomized graphs

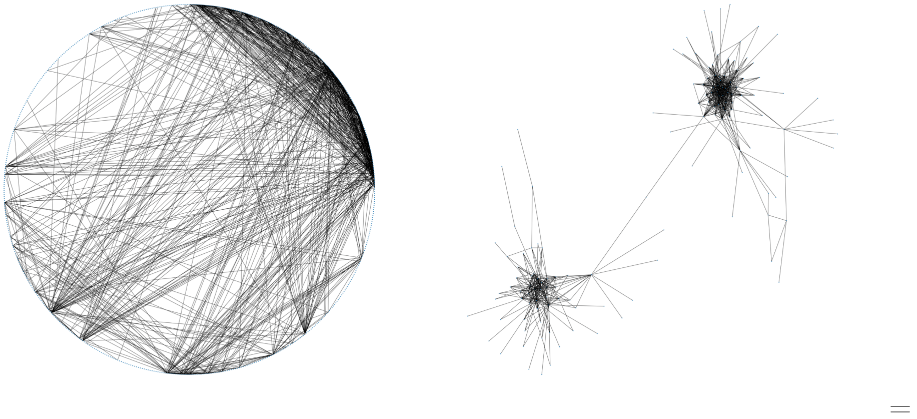
Using the Cytoscape `Motif_discovery` package, we searched for motifs on four taxa in the microbiota network graphs. This provided us with a count of the number of time that each motif appeared per disease cohort. We also used the Cytoscape `Network Randomizer` package to obtain a randomized graph.

We utilized the *Erdős–Rényi model*, the paradigm for generating randomized graphs using Bernoulli random variables, to validate experimental results and indicate a true correlation between sets of microbiota. We define a graph $G = (V, E)$ to be a tuple where $V$ is a nonempty finite set of vertices and $E$ is an edge set over $V$. Erdős–Rényi generates a random graph $G(n, m)$ as a function of number of edges and vertices $n = |V|, m = |E|$, and $G(n, m)$ thus had $\binom{n}{m}$ elements that each occur uniformly at random, i.e. with probability $1/\binom{n}{m}$.

## 3. Results

### 3.1. Network modeling and graph representations

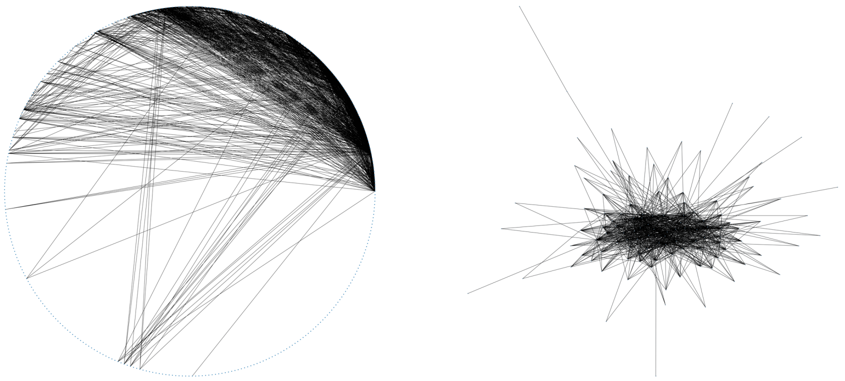
A first key step to network modeling was to threshold each cohort to see which taxa were correlated. Edges are drawn to indicate correlation on the family level where each plot is a network representation of the diseased cohorts.



a. IBS circular layout.

b. IBS clustering layout, no isolates.

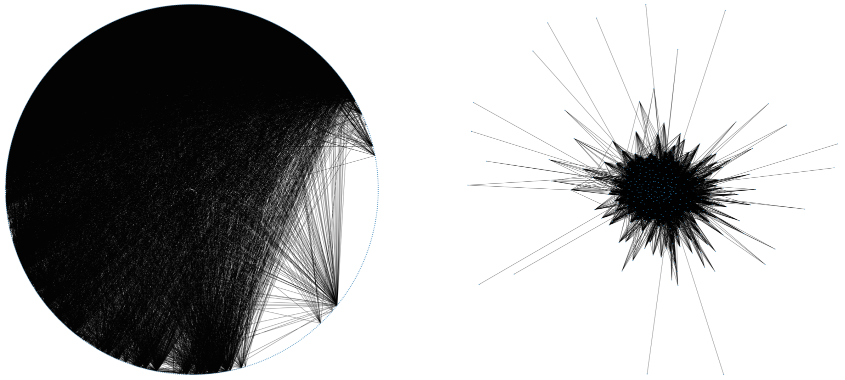
Figure 3. Network representation at the family level of the gut microbiome for irritable bowel syndrome, visualized using Python library `networkx` using both the circular heuristic and the regular clustering heuristic.



a. IBD circular layout.

b. IBD clustering layout, no isolates.

Figure 4. Network representation at the family level of the gut microbiome for inflammatory bowel disease, visualized using Python library `networkx` using both the circular heuristic and the regular clustering heuristic.



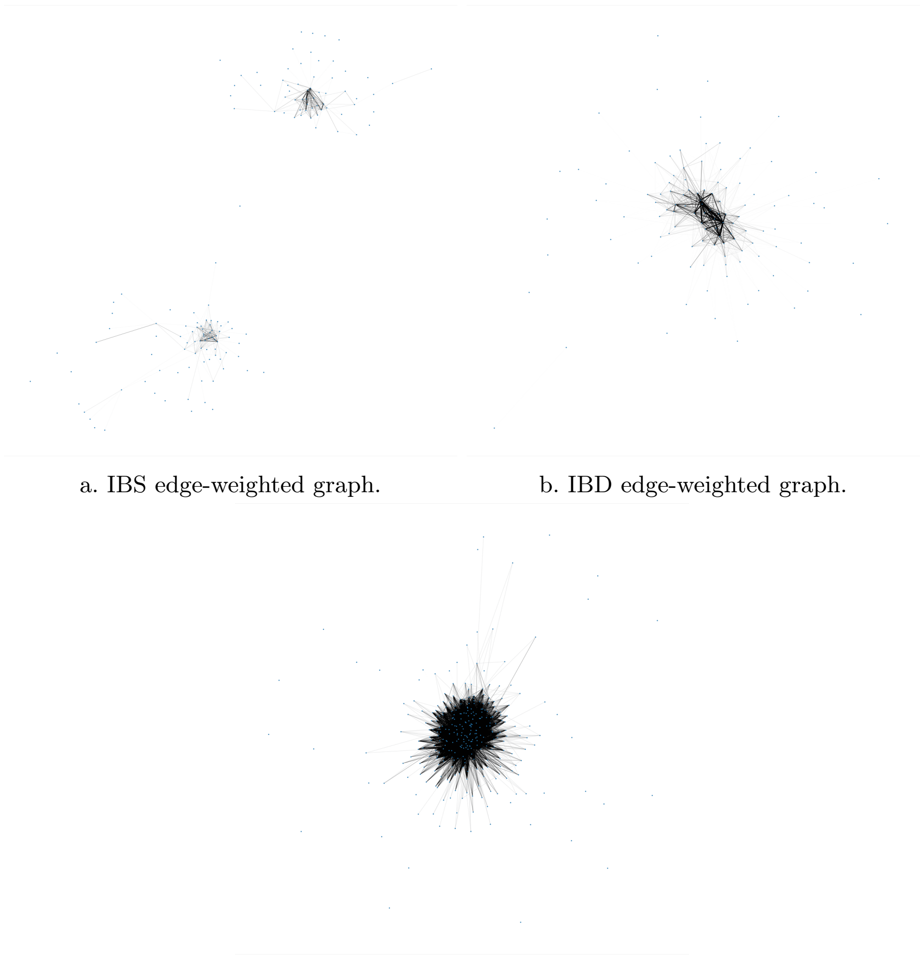
a. CRC circular layout.

b. CRC clustering layout, no isolates.

Figure 5. Network representation at the family level of the gut microbiome for colorectal cancer, visualized using Python library `networkx` using both the circular heuristic and the regular clustering heuristic.

### 3.1.1. Edge-weighted modeling

We present results from an analysis of edge-weighted models of the microbiota graphs of IBS, IBD, and CRC.



a. IBS edge-weighted graph.

b. IBD edge-weighted graph.

c. CRC edge-weighted graph.

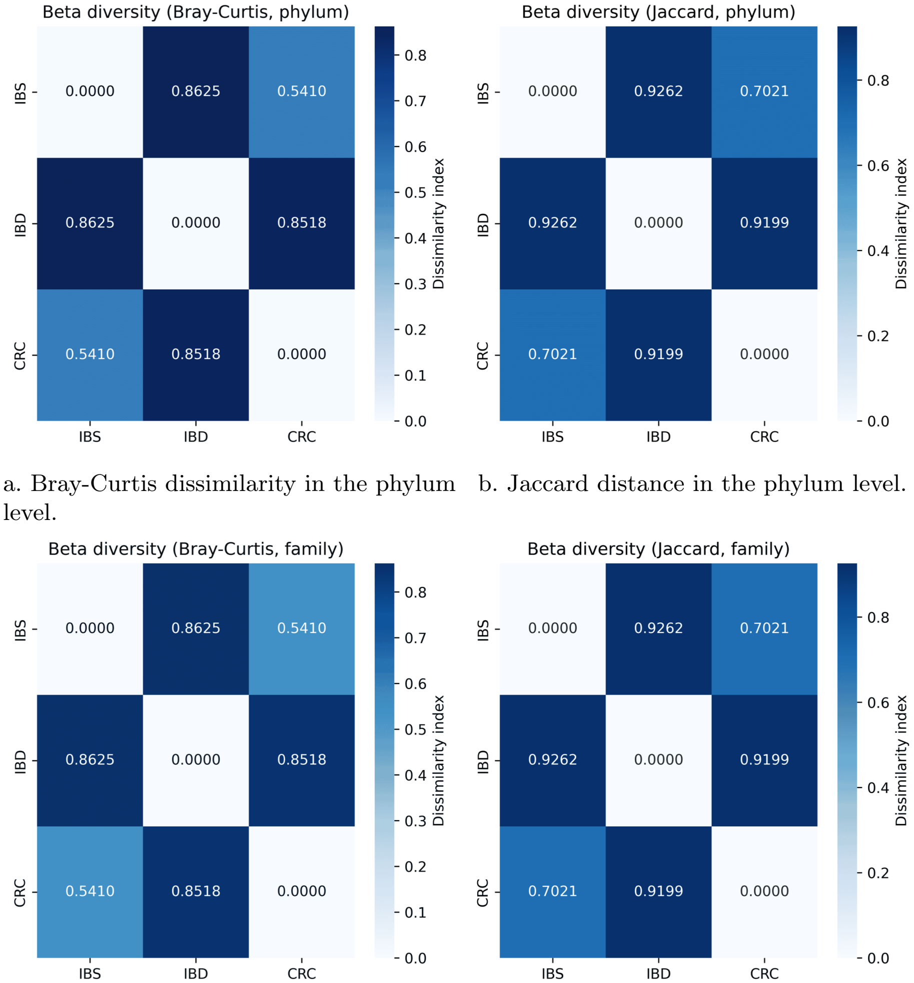
Figure 6. Network representation at the family level of the gut microbiome for IBS, IBD, and CRC, with normalized edge-weighting to indicate higher co-occurrence.

| | IBS | IBD | CRC |
|---|---|---|---|
| Average shortest unweighted path | 3.9655 | 1.9498 | 1.7491 |
| Average shortest weighted path | 2.6532 | 1.8426 | 1.1523 |
| Ratio | 0.6691 | 0.9450 | 0.6588 |
| Average weighted degree | 0.0490 | 0.0056 | 0.8979 |

Figure 7. Edge-weighted graph statistics, where a smaller ratio indicates a higher co-occurrence and correlation between the correlated taxa in the network. More correlated edges are given less weight to indicate a shorter distance. Average weighted degree in edge-weighted graphs without isolates, where more correlated edges are given more weight.

### 3.2. Core microbiota and biodiversity

Using the Bray-Curtis dissimilarity and Jaccard distance, we note the difference in diversity for the three cohorts, namely between IBD and IBS. These are measures of *beta diversity*, the ratio between regional and local species diversity. This was computed on both the order of phylum and family to illustrate how fine-grained these differences are.



a. Bray-Curtis dissimilarity in the phylum level.

b. Jaccard distance in the phylum level.

c. Bray-Curtis dissimilarity in the family level.

d. Jaccard distance in the family level.

Figure 8. Measures of beta biodiversity ranging from 0 (highly similar) to 1 (highly dissimilar) for IBS, IBD, and CRC.

We also present the alpha diversity of each disease cohort, calculated using a weighted generalized mean.

|  | IBS | IBD | CRC |
|---|---|---|---|
| Alpha diversity | 0.8878 | 0.6099 | 0.8424 |

Figure 9. Alpha diversity computed using Simpson's index calculated over all patients for IBS, IBD, and CRC. Lower values (0) represent low internal diversity while higher values (1) represent high diversity.

### 3.3. Motif search and network model topology

#### 3.3.1. Motif search and random graphs

We ran a motif search using Cytoscape to find subgraphs on four vertices common throughout the networks.

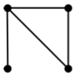
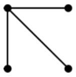
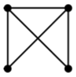
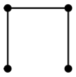
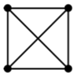
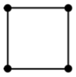
|  | **Irritable Bowel Syndrome**<br>(taxa: 139) | **Inflammatory Bowel Disease**<br>(taxa: 112) | **Colorectal Cancer**<br>(taxa: 313) |
|---|---|---|---|
|  | **0.2770%**<br>41257 total<br>Random: 28<br>0.00019% | **3.503%**<br>219262 total<br>Random: 165795<br>2.67% | **7.6789%**<br>30123634 total<br>Random: 21<br>0.000005% |
|  | **0.2291%**<br>34118 total<br>Random: 1342<br>0.009% | **2.9653%**<br>184171 total<br>Random: 131561<br>2.12% | **5.8256%**<br>22853356 total<br>Random: 1469<br>0.00037% |
|  | **0.1213%**<br>18062 total<br>Random: 0<br>0.0% | **1.2804%**<br>79525 total<br>Random: 64214<br>1.03% | **4.3774%**<br>17172101 total<br>Random: 0<br>0.0% |
|  | **0.1015%**<br>15109 total<br>Random: 4206<br>0.0282% | **1.3766%**<br>85498 total<br>Random: 65509<br>1.05% | **1.1931%**<br>4680247 total<br>Random: 4488<br>0.00114% |
|  | **0.0644%**<br>9583 total<br>Random: 0<br>0.0% | **0.5682%**<br>35287 total<br>Random: 27670<br>0.45% | **2.9587%**<br>11606895 total<br>Random: 0<br>0.0% |
|  | **0.0023%**<br>349 total<br>Random: 5<br>0.00003% | **0.0346%**<br>2148 total<br>Random: 1897<br>0.03% | **0.0279%**<br>109272 total<br>Random: 9<br>0.00000229% |

Figure 10. Motif search results on each disease cohort. Percentages are calculated using the number of total possible 4-cliques in the graph. The random graphs are the results from 1000 random graphs with the same number of edges and nodes as the original graphs.

### 3.3.2. Network model topology

We present a graph of the cumulative distribution function (CDF) of the degree of each taxa in each cohort, IBS, IBD, and CRC. Here, we observe a clear difference between the connectivity of each cohort, indicating potentially a larger number of interactions in the CRC cohort while the IBS and IBD cohorts both have relatively low numbers of interactions.

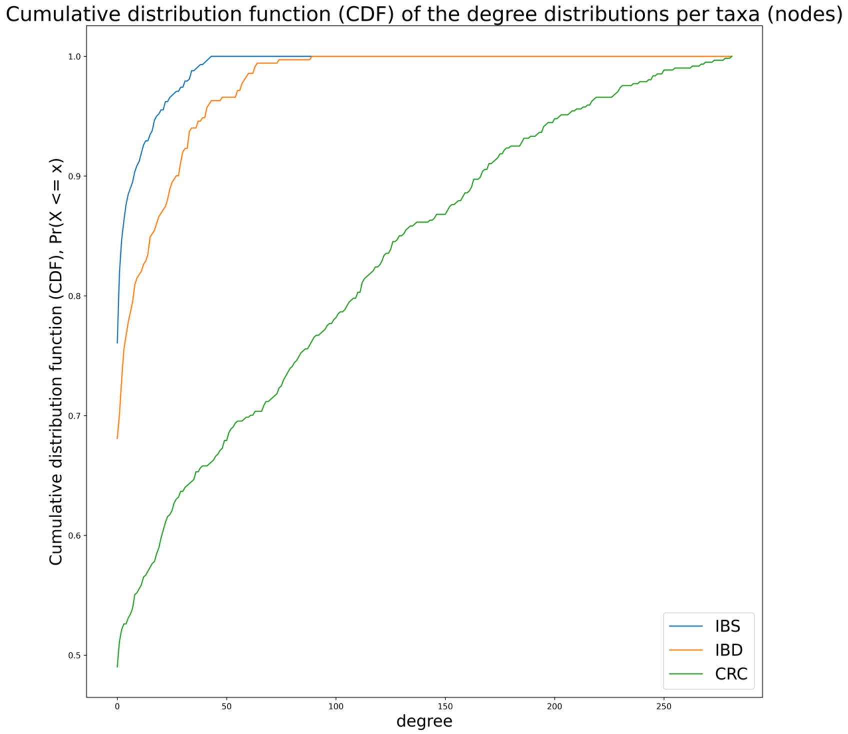


Figure 11. Cumulative distribution function of the degree of a vertex, i.e. $F_X(x) = \Pr(X \leq x)$, for disease classes IBS (blue), IBD (orange), and CRC (green). The $y$-axis representing the CDF values indicates a percentage.

This illustrates that CRC has a much higher degree of connected vertices than IBS and IBD, meaning it most likely has a richer and more interconnected microbiome. We also present the following microbiota graph properties:

|  | IBS | IBD | CRC |
|---|---|---|---|
| Number of vertices | 581 | 351 | 614 |
| Number of edges | 731 | 1149 | 14685 |
| Non-isolated vertices | 139 | 112 | 313 |
| Average degree | 1.2582 | 3.2735 | 23.9169 |
| Median degree | 0 | 0 | 0.5717 |
| Median non-zero degree | 3.02065 | 15 | 46.8762 |
| Clustering coefficient | 0.1361 | 0.2366 | 0.4164 |

Figure 12. General microbiota graph model statistics. The clustering coefficient is a measure of how much the nodes in a graph tend to cluster together, and represents a ratio of the actual number of edges to the total possible number of edges. All degrees are $\ell_1$-normalized with respect to the minimum number of vertices.

|  | IBS | IBD | CRC |
|---|---|---|---|
| Stability score | 0.8326 | 0.8215 | 0.6548 |

Figure 13. Stability score computed using the abundance-weighted mean interaction strength ($wMIS_i$ index).

These results indicate that the CRC microbiome is the most vulnerable to perturbations and can be easily changed, while the IBS and IBD ones have specialized such that it may be difficult to push them towards a more healthy state, or a different altered state. This signifies that using therapeutics to target microbiota in disease treatment may not be so simple.

## 4. Discussion

We identified taxa with high degree as key biomarkers of each disease, and we present the following list of top 5 biomarkers:

Irritable bowel syndrome:

1. *Isosphaeraceae*; degree 41
2. *Symbiobacteriaceae*; degree 38
3. *Oscillospiraceae*; degree 37
4. *Syntrophomonadaceae*; degree 34
5. *Archaeoglobaceae*; degree 34

Inflammatory bowel disease:

1. *Oscillospiraceae*; degree 89

2. *Lachnospiraceae*; degree 74
3. *Rikenellaceae*; degree 64
4. *Desulfovibrionaceae*; degree 63
5. *Peptostreptococcaceae*; degree 63

Colorectal cancer:

1. *Phycisphaeraceae*; degree 278
2. *Flavobacteriaceae*; degree 273
3. *Haloarculaceae*; degree 269
4. *Isosphaeraceae*; degree 267
5. *Gemmataceae*; degree 263

These findings are all supported by prior literature, and these families of microbiota are implicated in each of these disease cohorts. This means that these microbiota may serve as key therapeutic targets in future therapies.

### 4.1. Network Modeling

Figures 3, 4, 5 indicate that CRC is much more correlated than IBS and IBD, even after the $\ell_1$-normalization.

We define an *isolate* to be a vertex with no neighbors (i.e. $\{v \in V : N(v) = \varnothing\}$), and after filtering out isolates, it's clear in Figure 1 that the IBS microbiota cluster into two correlated regions, while IBD and CRC are one large region. This may be due to the two main types of IBS, IBS-C which is classified by hard stools and constipation, and IBS-D which is classified by loose stools and diarrhea.

In Figure 4, IBD has a dense network that seems to be shifted in terms of its bacterial populations from both IBS and CRC, suggesting that a specific network of bacteria have taken over the microbiome.

The CRC network is most dense with respect to the other networks in Figure 5, which suggest that their microbiota is less affected and is more robust for responding to changes in the environment. This aligns with experimental findings, as the gut microbiota in IBS and IBD are more clearly correlated with the two disease cohorts than the microbiome is with CRC. There is also an interesting outlier in the IBD cluster, corresponding to *Corynebacteriaceae*, a Gram-positive bacteria that assists with amino acid and enzyme production.

### 4.2. Edge-weighted modeling

Figure 6 indicates that the CRC graph is much more interconnected and taxa are more strongly co-correlated than in IBS and IBD. However, while the CRC network seems to be much more densely connected, the IBS taxa that do have neighbors or co-occuring taxa is similarly densely connected. With edges weighted from 0 to 1, where smaller numbers indicate more correlation, we have the following results in our graphs.

Since lower ratio numbers indicate more correlation, Figure 7 shows IBS and CRC have much stronger correlation among correlated taxa. IBD, however, seems to be very sparsely connected and not so correlated, as the edge weighted graph's shortest path is about as short as that of the unweighted graph. The average weighted degree also reflects these findings. CRC has significantly higher average degree than both IBS and IBD, but the average degree of IBS is still orders of magnitude higher than that of IBD. This truly indicates how sparse the microbiome of individuals suffering from irritable bowel disease is.

### 4.3. Core microbiota and biodiversity

Figure 8 displays trends in a higher beta diversity between IBS and IBD go across all levels of the phylogenetic tree, suggesting that while these two networks are highly interconnected and more rigorous, they are fundamentally different from each other. On the phylum level, the species appear to be much more similar, due to a lack of distinction between different microbiota. So beyond graph connectivity analysis, it's clear that there are key differences between IBS and IBD.

Furthermore, Figure 9 shows that the alpha diversity indicates that while the CRC network appears to be the most connected and robust, the IBS network actually contains a more diverse set of microbiota. This implies that gastrointestinal diseases do not strictly cause a loss in diversity, only a shift in interactions and the most expressed and dominant populations. IBD, however, does have a large loss in diversity, which aligns with current-day laboratory results.

### 4.4. Motif search and network model topology

Figure 10 indicates that CRC and IBD both have strongly connected sets of 4 taxa, while the microbiota in IBS are a lot strongly connected in groups. This suggests that the IBS network is more sparse, while the IBD network has a specialized network of core microbiota. It appears that the colorectal cancer network has more significant motifs as the network is just more well-connected.

Only the IBD random graphs have all six node connections present despite having fewer taxa, which suggests that that actual IBD nodes are much more connected compared to the CRC and IBS graphs. The CRC random graph have only 4 of the node connections despite having a much larger amount of nodes, therefore this microbiome is likely more diverse with less interactions between taxa. The IBS has similar connections to the CRC patients with much fewer nodes, so the graph is much more connected, but the microbiome is less diverse.

Figure 12 shows that the clustering coefficient, average degree, and median non-zero degree of the CRC cohort indicates a higher degree of connectivity. As this data is $\ell_1$-normalized, it is not biased by the number of vertices and this is an accurate metric. Furthermore, IBS has a disproportionately large amount of isolated vertices, leading to a lower average degree, indicating that it has a loss of abundance of potentially important gut flora.

Although the IBD cohort has the least raw correlation between its microbiota, in the normalized graph, those that are correlated are more correlated than the most correlated parts of IBS, implying that it has a small yet highly connected subset of gut microbiota.

Finally, we present the stability score of each network in Figure 13, which is an indicator of perturbation resistance (4). A more specialized network should have higher stability, while a more generalized network should be easily alterable to a diseased state. This is reflected in our findings.

### 4.5. Key implications

This network-modeling approach for representing the gut microbiome of two gastrointestinal diseases (irritable bowel syndrome and inflammatory bowel disease), as well as a gastrointestinal-related cancer (colorectal cancer) imply that the microbiota are key in disease progression and prognosis, and that they are correlated with each other and their co-occurrence can be indicative of certain disease classes and further useful as biomarkers and non-invasive diagnosis. These results have further implications as they present the loss of biodiversity in microbiomes of the gastrointestinal disease cohorts, implying that there is a loss of robustness. Furthermore, we are able to identify specific organisms that are significantly upregulated in disease classes, as well as those that are missing or significantly downregulated, providing a start for potential microbiota-targeted therapeutic treatments.

### 4.6. Comparison of methods and taxa choice

Microbiota-based graph construction can be done on all levels of the phylogenetic tree. We chose to construct a tree based on the family level due to the property that families are a group of genera with a common attribute, and using the genus scale would be much too fine-grained. However, running our graph-search algorithms on both order and phylum-based graphs yielded similar, although less detailed results. They indicated the same levels of changes in correlation but less significant information on the presence of motifs, the cumulative distribution function, and analysis of biodiversity.

### 4.7. Compute and data limitations

Running BLAST search on large quantities of 16S rRNA sequencing data is time and space intensive. Without access to large servers, we were constrained to only using reads from about 1/3 of the available patient data. The read sequences presented in `.fastq` files, although already compressed, takes up hundreds of GB of storage which is a resource that we did not possess. In the future, re-running our algorithms on a more robust subset of the available data may yield more accurate or statistically significant results.

This would also give us the ability to incorporate more modalities of data into our analysis, and analyze more diseased states beyond IBS, IBD, and CRC as we will be able to run BLAST search and store data from more patients. Furthermore, we notably did not possess high-resolution 16S rRNA-sequencing data for a healthy patient cohort. There was no 16S rRNA data found collected in a similar setting, and using incomplete or different data would create batch effect and provide unreliable results. In the future, once 16S rRNA-sequencing data is more largely available, an avenue of work could be to run these same network analysis algorithms on a cohort of healthy individuals.

### 4.8. Comparison with prior literature

Prevously, researchers compared patients with healthy microbiomes with patients suffering from and recovering from Alcoholic Liver Disease (ALD) using a similar network-based modeling approach (5). Using two specific sequences of 16S rRNA data, they constructed unweighted network graphs and ran motif searches to find patterns of interactions among the disease classes, and concluded that there was a disruption in the microbiota in diseased classes that was normal in healthy ones.

Our findings are more substantive and accurate as we consider all available 16S rRNA-seq reads from any given patient rather than just two. Our results align with theirs in the shifts in motif patterns and a massive loss of connectivity in IBD patients compared to IBS and

CRC patients. While we were unable to find sufficient healthy patient read data, looking at their results indicate that the healthy patient would likely have a more interconnected graph that differs greatly from the disease cohorts.

Another paper explores the relationships between host genes and gut microbiota using our same 16S rRNA-seq data and observed that certain species such as *Peptostreptococcaceae* were associated with host gene pathways in IBD (6). We observe this in our results as well, as in IBD, *Peptostreptococcaceae* co-occurs with 63 other taxa, which is approximately 20 times the average degree. This indicates that our microbiota networks are representative of true biological results. We discover similar results in the implication of alterations in human gut microbiota in gastrointestinal conditions.

### 4.9. Conclusions and future work

In this paper, we used network-modeling to represent the gut microbiome for two gastrointestinal diseases (IBS and IBD) and gastrointestinal-related cancer (CRC). Our results imply that the microbiota are key in disease progression and prognosis.

The data has demonstrated that a key characteristic of the IBS microbiome is the lack of diversity, in IBD the overgrowth of pathogenic bacteria, and in CRC loss of beneficial bacteria. These are all forms of dysbiosis, which refers to the ways in which the microbiome can be imbalanced. Thus, dysbiosis of the microbiome commonly causes inflammation which can exacerbate the disease and also lead to other gastrointestinal diseases, the way IBD is a risk factor for CRC. This demonstrates that it is important to prevent inflammation from occuring in the first place. Inflammation is caused by the bacteria breaking down the lining of the GIT, however, if they have other food sources, such as those from a diverse diet, this will help mitigate the amount that bacteria break down the lining.

Studies note that incorporating metatranscriptomic and metabolomic sequencing is necessary for a more robust understanding of the differences in the gut microbiota. There is a large variety of data for the human gut microbiome, and while we were only able to focus in 16S rRNA-sequencing data, further avenues of work could incorporate other modalities of data to construct a more accurate network (7). Constructing a biochemical knowledge-driven (bottom up) microbiota network using genome annotation and known chemical reactions could lead to more robust differentiation analysis as well as disease cohort classification. Further analysis could be done on microbiota of patients throughout the progression of a certain gastrointestinal disease, and observe the shifts in co-occurrence and connectivity over time.

### 4.10. Data sources and availability

The high-resolution raw data for the 16S rRNA sequencing (rRNA-seq) for the CRC cohort can be accessed at PRJNA2843553. The 16S rRNA-seq reads for the IBD cohort can be accessed at PRJNA398089 and the 16S rRNA-seq reads for the IBS cohort can be accessed at PRJEB37924. These reads were sequenced using metagenomic assembly by the University of Minnesota and European Nucleotide Archive (ENA).

### 4.11. Code availability

Code used for all topological analyses performed in the paper is available at the repository ARRiVAL. The installation instructions for `RDPTools` and setting up and running the RDP Classifier can be found here. Software used for graph analysis can be found at GraphCrunch and Cytoscape.

# References

[1] Matthew J. Bull and Nigel T. Plummer. Part 1: The human gut microbiome in health and disease. *Integrative medicine*, 13.6, 2014.

[2] The Integrative HMP (iHMP) Research Network Consortium. The integrative human microbiome project. *Nature*, 569, 2019.

[3] J. Michael Janda and Sharon L. Abbott. 16s rrna gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and pitfalls. *Journal of Clinical Microbiology*, 45.9, 2007.

[4] Yinhu Li, Yijing Chen, Yingying Fan, Yuewen Chen, and Yu Chen. Dynamic network modeling of gut microbiota during alzheimer's disease progression in mice. *Gut Microbiome*, 15.1, 2023.

[5] Ammar Naqvi, Huzefa Rangwala, Ali Keshavarzian, and Patrick Gillevet. Network-based modeling of the human gut microbiome. *Chemistry & Biodiversity*, 7.5, 2010.

[6] Sambhawa Priya, Michael B. Burns, Tonya Ward, Ruben A. T. Mars, Beth Adamowicz, Eric F. Lock, Purna C. Kashyap, Dan Knights, and Ran Blekhman. Identification of shared and disease-specific host gene–microbiome associations across human diseases using multi-omic integration. *Nature Microbiology*, 7, 2022.

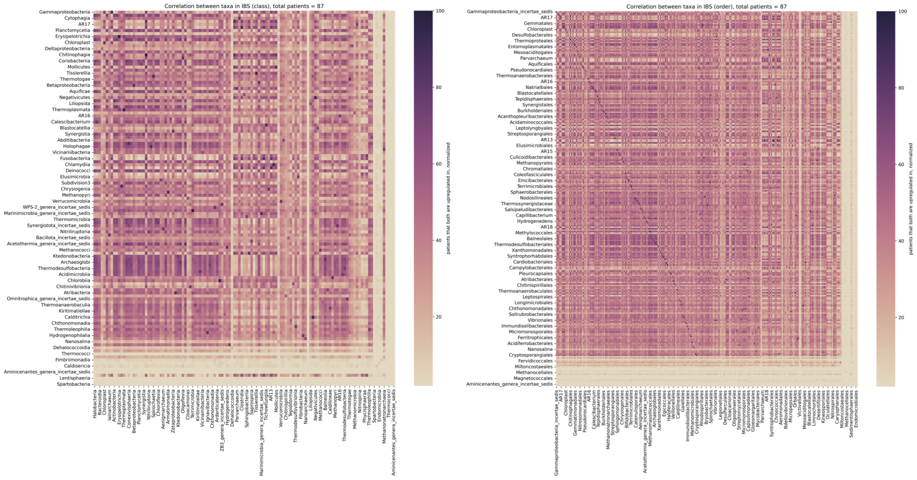
[7] Sofia D. Shaikh, Natalie Sun, Andrew Canakis, William Y. Park, and Horst Christian Weber. Irritable bowel syndrome and the gut microbiome: A comprehensive review. *Journal of Clinical Medicine*, 12.7, 2023.

[8] Qiong Wang, George M. Garrity, James M. Tiedje, and James R. Cole. Naïve bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, 73.16, 2007.

[9] Chi Chun Wong and Jun Yu. Gut microbiota in colorectal cancer development and therapy. *Nature Reviews Clinical Oncology*, 20, 2023.

# 5. Appendix

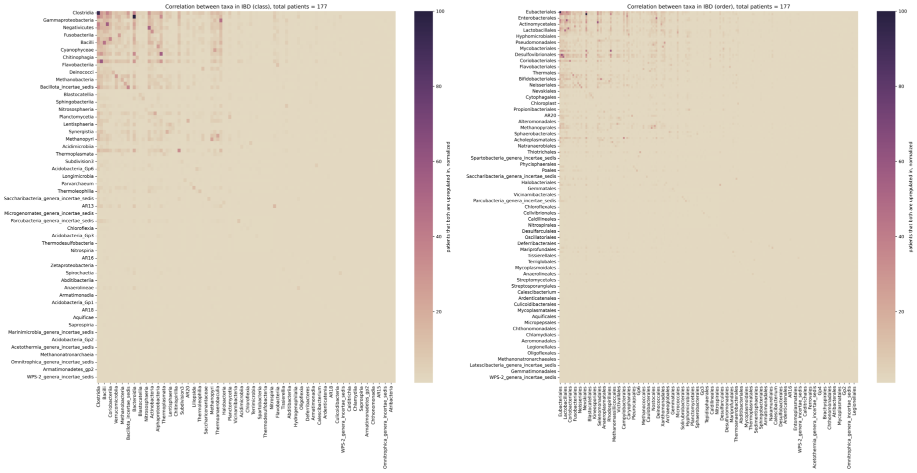
## 5.1. Taxa co-occurrence heatmaps



a. IBS class.

b. IBS order.

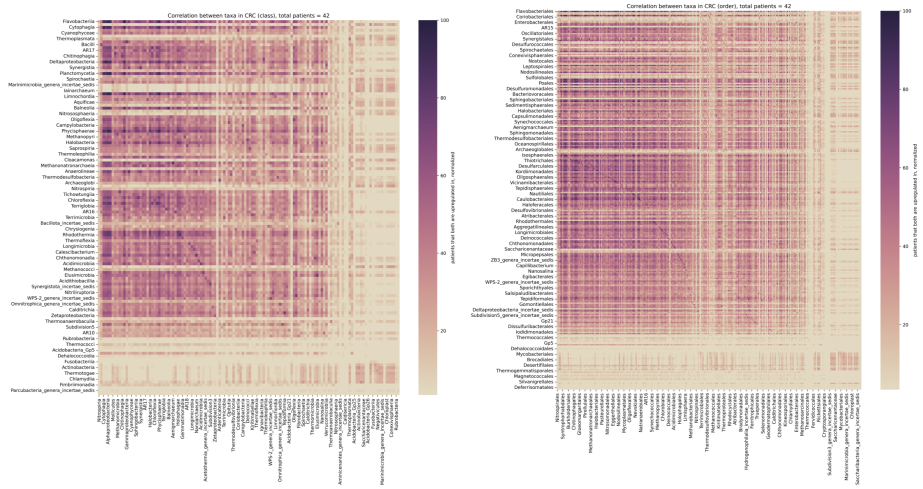
Figure A1. Measures of taxa co-occurrence for IBS on the levels of class and order.



a. IBD class.

b. IBD order.

Figure A2. Measures of taxa co-occurrence for IBD on the levels of class and order.

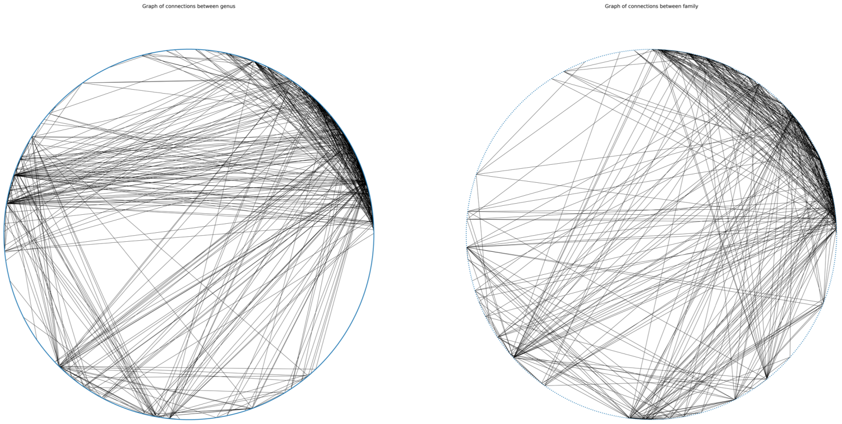


a. CRC class.

b. CRC order.

Figure A3. Measures of taxa co-occurrence for CRC on the levels of class and order.

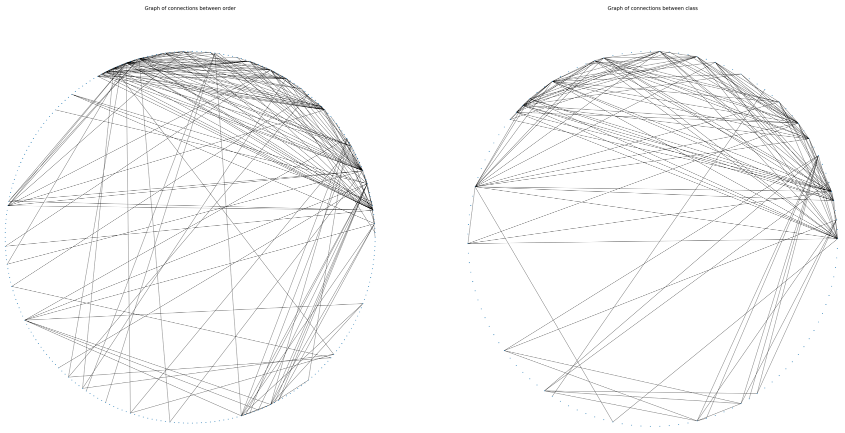
## 5.2 IBS graphs at all hierarchical levels using absolute clustering


Graph of connections between genus
Graph of connections between family

a. IBS genus co-occurrence graph.

b. IBS family co-occurrence graph.

Figure A4. Graphs of taxa co-occurrence for IBS on the levels of genus and family (0.4 absolute). Values correspond with what percentage of the maximum of the patients both are upregulated in.


Graph of connections between order
Graph of connections between class

a. IBS order co-occurrence graph. b. IBS class co-occurrence graph.

Figure A5. Graphs of taxa co-occurrence for IBS on the levels of order and class (0.4 absolute). Values correspond with what percentage of the maximum of the patients both are upregulated in.

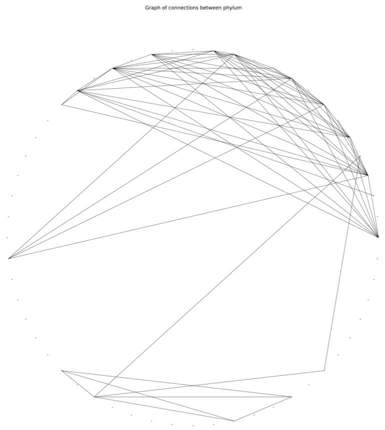


Figure A6. Graphs of taxa co-occurrence for IBS on the level of phylum (0.4 absolute). Values correspond with what percentage of the maximum of the patients both are upregulated in.