“Meta-analytic investigation of gut microbial community structure identifies a panel of stability-promoting microbiome members consistently reduced with gut inflammation”

by
Amit Samal

Under the supervision of
Dr. Tarini Shankar Ghosh

Submitted in partial fulfillment of the requirements for the degree of Master of Technology, Computational Biology

Center for Computational Biology, Indraprastha Institute of Information Technology - Delhi
August, 2023

Certificate

This is to certify that the thesis titled *"Meta-analytic investigation of gut microbial community structure identifies a panel of stability-promoting microbiome members consistently reduced with gut inflammation"* being submitted by **Amit Samal** to the Indraprastha Institute of Information Technology Delhi, for the award of the Master of Technology, is an original research work carried out by him under my supervision. In my opinion, the the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

August, 2023

Dr. Tarini Shankar Ghosh

Department of Computational Biology
Indraprastha Institute of Information Technology Delhi
New Delhi 110 020

Acknowledgements

I sincerely thank Prof. Tarini Shankar Ghosh for his unwavering guidance and mentorship throughout this journey. Without his support, this project would not have been possible. From the beginning of my thesis, he continuously challenged me to reach higher goals and excel in my work. He generously offered his time to address any issues or clarify my doubts, even during busy days. Whenever I encountered challenges in my work, he provided valuable insights and refined our approach, significantly contributing to the timely completion of my tasks. I am truly grateful for his constant encouragement and support, which have been instrumental in my academic and personal growth.

I would like to express my sincere gratitude to my parents and sister for their unwavering support, which has been invaluable throughout my work. I am thankful for their outstanding commitment, which has played a vital role in shaping my academic journey.

Throughout my thesis work, I turned to my friend Anshul Yadav for any small or trivial doubts related to my research or career advice, and I am immensely grateful for his support.

I also thank my lab mates, Abhishek, Lavanya, and Shivansh, for supporting me throughout the work. Their presence and encouragement were invaluable.

I extend my heartfelt gratitude to all the faculty members and staff of the Department of Computational Biology at IIIT Delhi for their constant support and assistance throughout my college journey.

Abstract

Inflammatory Bowel Diseases (IBD), encompassing Crohn's Disease (CD) and Intestinal Tuberculosis (ITB), present similar clinical symptoms but require distinct treatment strategies. This study investigates microbiome and mycobiome alterations in both diseases (ITB & CD) and compares the dysbiosis concerning the controls. The study also investigates the critical role of non-bacterial components, like fungi, in differentiating both diseases. The level of dysbiosis is dissected through a network of co-abundant modules for a set of diagnostic biomarkers and validated globally using ¿5,400 gut profiles. The central role of disease-specific depleted core microbiomes was verified by analyzing the reproducibility of these taxa in depleted groups across many available disease data sets. The Core Indian gut microbiomes are identified to get the health status of a disease subject. Subsequent investigation of these markers across greater than 5,000 longitudinal gut microbiomes from 12 diverse studies reveals consistently strong positive associations between the abundance of these markers and the long-term stability of the gut microbiome. Our study, for the first time, identifies and highlights the role of specific central taxa as putative protectors against gut inflammation and in promoting the long-term stability of the gut microbiome. These findings can potentially advance the development of specific microbial consortia pro-biotic supplementation addressing inflammation-associated gut dysbiosis.

Keywords: Inflammatory Bowel Diseases, Microbiome, Mycobiome, Co-abundance Networks, Disease-specific Dysbiosis, Longitudinal Resiliency.

Contents

List of Figures

Chapter 1

Introduction

Inflammatory bowel disease (IBD) is a condition with chronic inflammation in the gastrointestinal tract. Alteration in the interaction between the host intestinal immune system and microbial community in the gut results in such a critical health condition. Microbial communities in the gut maintain the gut barrier and influence the production of mucus and peptides with antimicrobial activities, which protects the gut from pathogenic colonization. They also help in energy balance, host nutrient metabolism, and immune regulation by producing short-chain fatty acids (SCFAs) such as butyrate, acetate, and propionate by breaking down dietary fiber [1].

The gut immune system interacts with microbiome communities by recognizing and responding to the microbial antigen. This also helps the immune system to maintain immunotolerance to commensal microbiota [2]. Most of the gut microbiota live within the host by maintaining a relationship of mutualism wherein some symbionts like *Faecalibacterium prausnitzii* support host homeostasis by producing butyrate, an essential SCFA. In contrast, there are other microbial lineages like *Fusobacterium nucleatum* that, in certain conditions, can exert putative detrimental effects by producing proinflammatory chemokines interleukin, IL-6, and IL-8[3]. Microbial modulation of host immunity is still a hitherto unresolved aspect of host-microbiome interactions, where the exact mechanisms of how certain microbial species modulate host immune responses to maintain their survival and growth in the host gut are largely unknown. The composition of these mutualistic microbiome communities also varies among individuals and in the same individuals across different periods [4].

The two most prevalent forms of IBD are ulcerative colitis (UC) and Crohn's disease (CD), with inflammation characterized by multiple relapses after remission[5]. Alteration in the host immune response and changes in the gut microbial community caused due to the host response as well as due to many environmental factors (including

diet, medications, and social exposure), are closely associated with the onset of many gut-inflammation-associated diseases. Patients with these diseases show diarrhea, abdominal pain, rectal bleeding, and weight loss symptoms[5]. The dynamics and state of the microbial communities in a healthy individual are mostly personalized. Many of the prior studies have revealed a state of dysbiosis resulting in disruption of the composition of the gut microbes and their functionality in most cases of IBD patients[6].

Tuberculosis (TB) is a deadly infectious disease caused by *Mycobacterium tuberculosis*. The primary infection site of *M. tuberculosis* is mainly in the pulmonary region, but it can infect almost any part of the human body[7]. TB can present in a dynamic condition in the human body, from an asymptomatic state to a very contagious life-threatening disease [7, 8]. Tuberculosis patients generally experience multiple symptoms, like fever, lack of appetite, and in many cases, weight loss. Intestinal tuberculosis is one of the major forms of these extra-pulmonary manifestations of TB[9]. The gastrointestinal tract ranks sixth most preferable site of infection by M. tuberculosis [9]. In the past two decades, many drug-resistant TB strains have been reported, increasing ITB incidence in both developing and developed countries [10]. Although the causes of ITB and CD differ, they share similar clinical, pathological, radiological, and endoscopic features[11]. The distinction between these two diseases is very challenging, especially in regions like India, where TB is endemic, and the number of IBD cases is increasing. The increasing burden of Inflammatory bowel Diseases (IBD) in the developing world has been attributed to dietary and lifestyle changes because of rapid modernization and Westernization of the population[12]. In the laboratory setup, a positive acid-fast bacillus (AFB) culture, Caseation necrosis on biopsy samples, and computed tomography (CT) images of necrotic lymph nodes are the few ways to differentiate the ITB infections from the CD with noticeably high specificity but very poor sensitivity[11]. The strategies implemented in the therapy between ITB and CD are also very different. The former is treated with anti-tuberculosis drugs, whereas CD patients mostly use immune-suppressive agents. The similar clinical manifestation between these two diseases leads to misdiagnosis, alternately worsening the condition of patients who miss the best time for treatment[13].

There are several changes in the microbiome composition and diversity in IBD patients compared to healthy individuals. So the stratification of the microbial communities by their loss or gain to differentiate among CD, ITB, and healthy controls has been a focus of active research[14]. Another component that requires comprehensive investigation is the fungal component in the gut micro-environment (referred to as the mycobiome). There are several studies in which CD-associated microbiome and mycobiome profiles are available now, but the studies on bacterial and fungal profiles for intestinal tuberculosis are very limited. The importance of gut bacteria in disease pro-

gression and homeostasis is reasonably known, and they are the most abundant player in the gut microenvironment[15]. In the last two decades, most studies focused on bacterial composition and its effect on human health. However, recently, the role of the dynamics and interactions of the non-bacterial components of the microbiome in the onset of different diseases have been revealed, pointing out their critical role (as well) in host well-being [15]. Even though fungi are present in very small proportions in the gut microbial communities, their size and metabolic potential make them significant contributors to gut physiology.

The current study aims to identify the distinguishing microbiome markers of CD and ITB and their variations with respect to apparently non-diseased control individuals and then investigate the association of specific microbial marker subsets identified with the long-term stability of the gut microbiome abundant communities. The first part of the current study focused on differentiating the change in composition and diversity of the gut bacteriome/archaeome and mycobiome associated with ITB from that of CD-associated communities(Figure 1.1).

With information related to the community-level dysbiosis in the two diseases, we next identified the microbiome and mycobiome signatures specific to each disease type and their common signatures (with respect to non-diseased controls). In analyzing the disease-specific dysbiosis in the microbiome and mycobiome, we examined various intra-community level co-abundant-module in both diseases and their reproducibility across multiple cohorts with depleted and enriched abundances. We also validated the reproducibility and clinical significance of these modules by performing a large-scale global analysis of intra-microbiome, intra-mycobiome, and microbiome-mycobiome meta-networks with 5,400 gut bacteriome profiles and 900 gut mycobiome profiles collated from 14 studies across the globe. We analyzed the gut microbiome profiles for three sample types – CD, ITB, and controls from this study, along with 1617 gut microbiomes from seven other studies to identify the set of core Indian gut microbiomes. In this core gut microbiome profile, 28 species out of the 46 members of the microbial co-abundance depleted modules across the ITB and CD profiles hold a central position with greater than 70 percent prevalence. Finally, the core gut microbiome profiling resulted in the formulation of a method to quantify the retention of core species in an individual.

In the second part, the study has been extended to the dysbiosis of microbiomes in the gut from a single observation time to a longitudinal time point analysis of the microbial communities focusing on accounting for the diversity at multiple observational

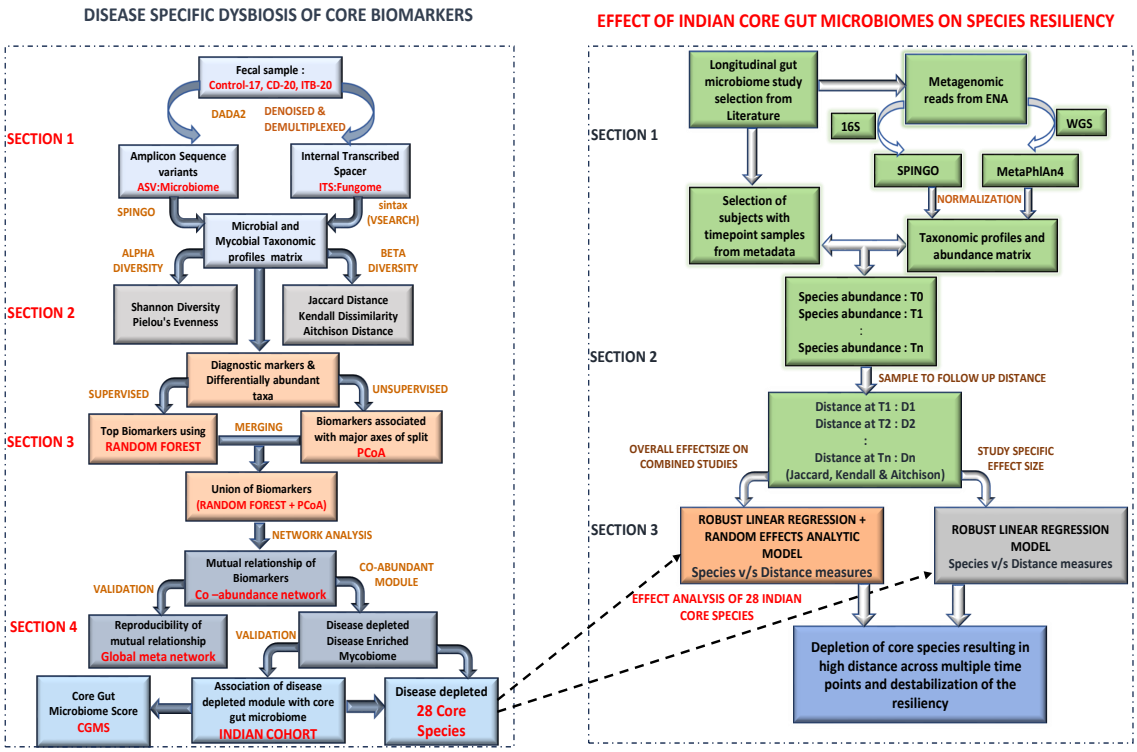

Figure 1.1: Process flow explaining the two stages of the study

landscapes. This longitudinal study aimed to analyze the resiliency in the microbiome abundances across time points and assess the core groups' effect in the depletion or breakage of co-abundance networks in disease cases. A complete 12 independent multi-time-point studies of multiple gut-related diseases and controls across various countries were utilized to avoid any region-specific bias in this longitudinal resiliency study. The validation of these core microbiomes in our longitudinal studies showed >70% reproducibility of these species with low negative effects on the beta diversity across multiple time points. The overall effect of these depleted microbiomes across the studies gave us an idea of the central position they play in maintaining the chain of co-abundance in a healthy gut, and the depletion of the global core resulted in lesser beta diversity across multiple observation time points.

Chapter 2

MATERIALS AND METHODS

2.1 Data set collection and processing

The disease group samples (ITB/CD) and control group samples in our current study were collected at the All India Institute of Medical Sciences, New Delhi. A sum total of 57 samples were included in the study. In both ITB and CD,20 samples were included, whereas in control cases, 17 samples were included. The raw 16S and ITS (Internal Transcribed Spacer sequence) sequence reads from AIIMS were shared with Microbiome Informatics Lab at IIIT Delhi(https://microbiome.iiitd.edu.in/). The remaining data sets used in the validation and the core gut microbiome score (CGMS) calculation can be found in the table[Table 2.1]. The raw sequence reads, except for AIIMS samples, were downloaded from the European Nucleotide Archive(ENA)(https://www.ebi.ac.uk/ena/browser/home). Then the downstream processing, statistical, and machine learning-based analysis were done at IIIT Delhi.

The data sets in resiliency studies were shortlisted after literature reviews of gut-related infection of patients from many publicly available databases [Table 2.2]. All the data sets were verified for the availability of the longitudinal subject(individual patient) data points. The subjects with multiple time point samples were selected, and the remaining were excluded from the study. The raw sequence reads of 16S and whole genome sequences was downloaded from the European Nucleotide Archive(ENA) (https://www.ebi.ac.uk/ena/browser/home). As mentioned above, the sequenced data were processed and analyzed in various statistical and machine-learning tools for further inference.

Table 2.1: Data sets used for co-abundant network validation and identification of Core Indian gut microbiomes

Dataset	Datatype	Total Sample		Nationality	Purpose		
		Microbiome	Mycobiome		Meta Network analysis	Indian Core Gut Microbiome	CGMS validation
AIIMS 2021[16, 17]	16S	162	0	India	Yes	Yes	Yes
MicroDiab India[18, 19]	16S	435	0	India	Yes	Yes	No
LogMPie[20]	16S	874	0	India	Yes	Yes	No
Dhakan DB 2019[21]	Shotgun	88	0	India	Yes	Yes	No
Gupta A 2019[22]	Shotgun	60	0	India	Yes	Yes	No
PRJDB7616[23]	16S/ITS	58	58	India	Yes	Yes	No
	16S/ITS	39	39	Japan	Yes	Yes	No
Liguori 2016[24]	16S/ITS	47	47	Italy	Yes	No	Yes
PRJEB423575[25, 26]	16S/ITS	115	144	Ireland	Yes	No	No
HMP2[27]	Shotgun/ITS	0	390	US	Yes	No	No
PRJNA439151[28]	ITS	0	70	China	Yes	No	No
PRJNA647266[29]	ITS	0	75	China	Yes	No	No
He et al[30]	16S	3559	0	China	Yes	No	No
PRJNA662173[31]	ITS	0	69	India	Yes	No	No
Halfvarson 2017[6]	16S	683	0	Sweden	No	No	Yes
Lloyd-Price 2019[5]	Shotgun/ 16S	1627	0	USA	No	No	Yes
Hall 2017[32]	Shotgun	259	0	USA	No	No	Yes
Franzosa 2018[2]	Shotgun	219	0	USA	No	No	Yes

Table 2.2: Data sets used for resiliency study

Study Name	Country	Condition	Total Sample	Total Subject
Halfvarson_2017[6]	Sweden	Collagenous colitis(CC), CD, Control(HC), Lymphocytic Colitis(LC), UC	541	118
Wedenoja_2022[33]	Finland	Chronic Liver Disease(CLD)	117	28
Mars_2020[34]	USA	Control, IBS	434	71
LloydPrice_2019[5]	USA	Control, UC,CD	1491	111
Palleja_2018[35]	Denmark	Control	45	12
Kang_2022[36]	Singapore	Carbapenemase-producing Enterobacteriaceae (CPE positive), CPE Negative, Non CPE colonized family member	311	41
Clooney_2020[37]	Ireland/Canada	CD, Coeliac, UC,IBS, Control	969	549
Tee_2022[38]	Malaysian	Helminth Positive, Control	243	132
Metwaly_et_al[39]	Spain	CD	115	14
NUAGE[40]	Europe	Control	614	614
RaymondF_2016[41]	CANADA	(Control, Cephalosporin treated)	48	24
VincentC_2016[42]	CANADA	Healthy Control, Abdominalhernia, Clostridium difficile colitis, Cellulitis, Pneumonia, Ureteralstone, Gangrene	129	69

2.2 Methods

2.2.1 Computation of Microbiome (Bacteria & Archaea) and Mycobiome (Fungi) abundance profiles

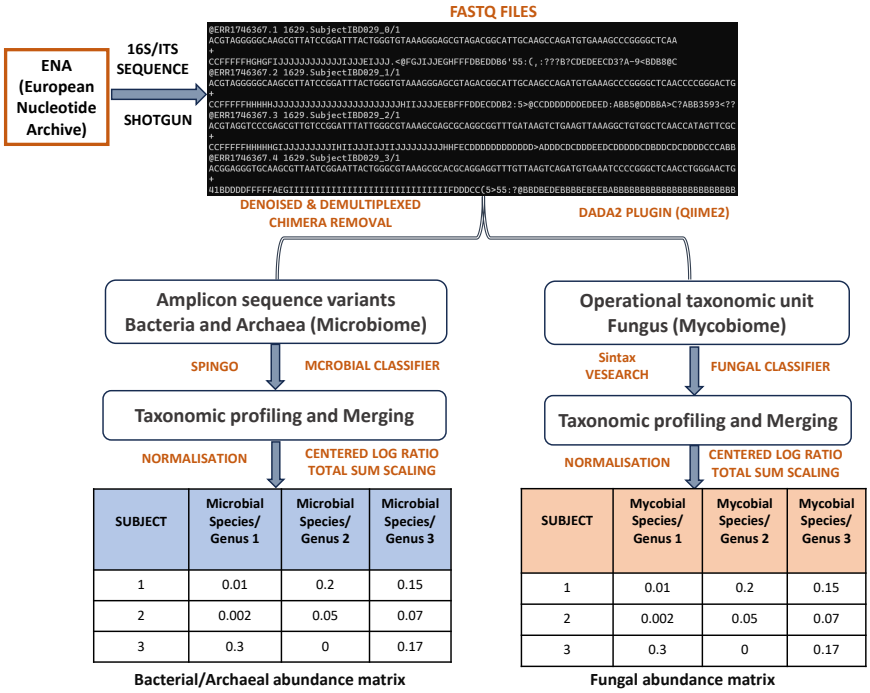

Figure 2.1: Data processing pipeline for the raw sequences received from AIIMS, Delhi

The raw sequence reads were analyzed using QIIME2 version 2022.2[43]. The forward and reverse paired-end reads were processed, denoised, and demultiplexed. Additionally, chimera removal was performed using the DADA2 plugin of QIIME2[44]. The DADA2 plugin was also used to generate Amplicon Sequence Variants (ASVs) for the 16S (for Bacteriome and Archaeome) and ITS (Internal Transcribed Spacer) for Mycobiome data sets, enabling the detection profiles of each ASV in each of the samples [Figure2.1]. The representative sequences of each 16S-derived ASV (Bacteriome/Archaeome) were classified at the Species and Genus levels using the SPINGO classification tool[45]. Similarly, for the ITS data, the classification was performed using the 'sintax' taxonomy classification tool within the 'vsearch' annotation pipeline, with the UNITE database as the reference[46]. With the obtained information, Species and

Genera abundance profiles were obtained separately for 16S and ITS data by accumulating the abundance profiles of ASVs assigned to each Species and Genus, respectively.

2.2.2 Investigating the overall alpha and beta-diversity variations across the three groups of gut microbiomes

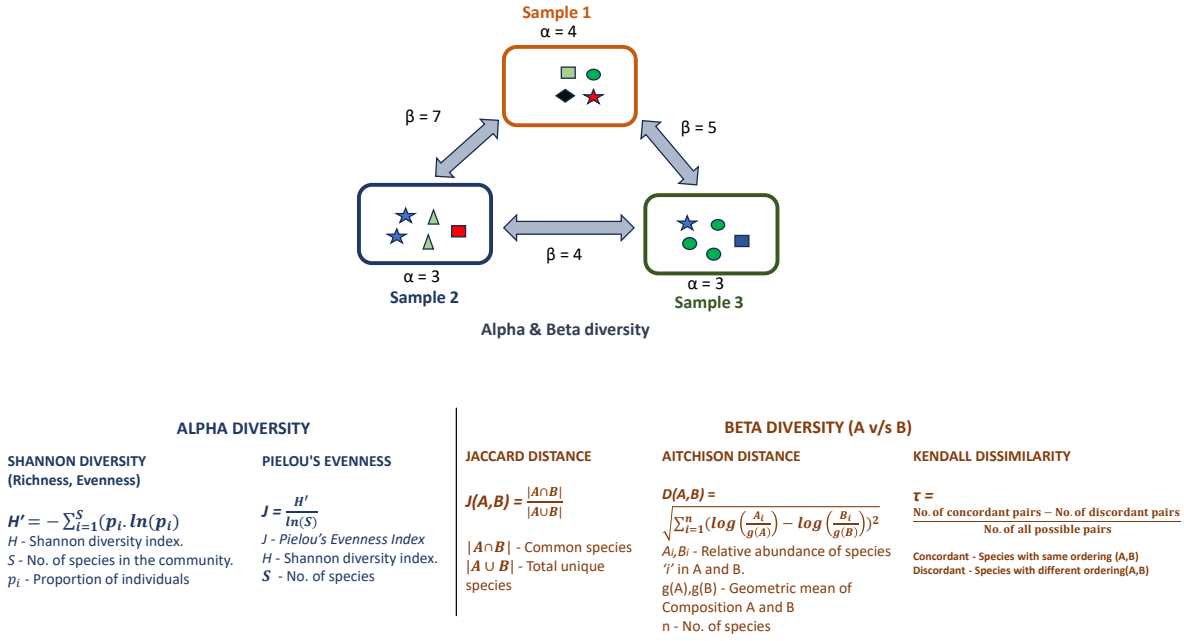

Figure 2.2: Alpha and Beta diversity measures

Both alpha and beta diversity measures are explained in Figure[2.2]. The diversity and richness measures, including the 'Shannon Index' and 'Pielou Evenness', were calculated using the R package vegan (version 1.00.15). The Shannon Index was computed using the 'diversity' function. To calculate the Pielou Evenness index, we divided the Shannon Index values obtained from the 'diversity' function by the logarithm of the total number of taxa detected in each sample, calculated using the 'specnumber' function of the vegan package.

The 'Shannon' and 'Pielou Evenness' indices were calculated at the species and genus levels for the three subject groups' microbiome and mycobiome data sets. We used Kruskal-Wallis H-tests (performed with the 'kruskal.test' function in the R pro-

gramming interface) to compare the indices across the three groups. P-values indicating significant value variations between groups were computed using Dunn's tests (performed with the 'dunn.test' function in R) for pairwise comparisons. The resulting p-values were corrected using the Benjamini-Hochberg method by setting the 'method' argument to 'bh'.

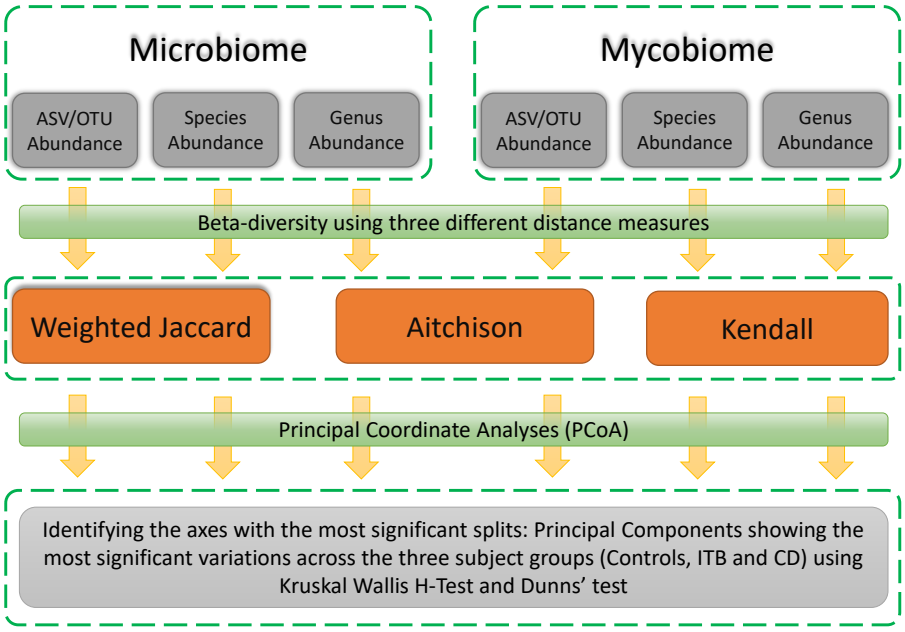

Figure 2.3: Beta diversity calculation method

The beta diversity analysis across the samples was conducted using three different distance measures: Kendall, Weighted-Jaccard, and Aitchison [Figure 2.3]. The abundance profiles were normalized using the Total Sum Scaling (TSS) approach for the Kendall and Weighted-Jaccard distance measures. This involved dividing the counts of each taxon in each sample by the total sum of all taxa in that sample. On the other hand, for the Aitchison distance measure, the normalization was performed using the Centered-Log-Ratio (CLR) transformation approach. The CLR transformation aims to handle compositional data. After applying the CLR transformation to the abundance profiles, the Aitchison distances were computed as the Euclidean distances between samples based on these CLR-transformed abundance profiles. We first used the 'clr' function of the 'compositions' package in R v2.0.4 to execute the normalization, and then, we computed the Euclidean distances between samples using the 'vegdist' func-

tion, setting the 'method' argument to 'Euclidean' like below:

```
Aitchison Distance Matrix=
as.matrix(vegdist(clr_transformed_abundance_matrix, method = "euclidean"))
```

For computing Kendall distances, we utilized the 'cor.fk' function from the 'pcaPP' package v2.0.2. This function allowed us to generate the Kendall tau correlation matrix across the samples. Subsequently, we converted the Kendall correlations into a Kendall distance matrix using the following method:

```
Kendall Distance Matrix = as.matrix(1-cor.fk(t(abundance_matrix))/2)
```

Finally, the weighted Jaccard distances were computed as:

```
Jaccard Distance Matrix =
as.matrix((vegdist(abundance_matrix, method="jaccard"))
```

Each distance matrix explains different aspects of variations in the microbiome using different transformations [47]. Separate matrices were used for each distance measure for microbiome and mycobiome profiles.

The beta-diversity variations for each distance measure and profile type (microbiome and mycobiome) were visualized and analyzed using Principal Coordinate Analysis (PCoA). PCoA was conducted using the 'ade4' package v.1.7.17 of R. Specifically, six separate PCoA analyses were performed, corresponding to the three distance measures (Kendall, Aitchison, and Weighted Jaccard) applied to both the microbiome and mycobiome profiles.[Figure 3.2,3.3,3.4]. In each of the six PCoA plots, we visually investigated the separation between the three subject groups. Next, we compared the major PCoAs that exhibited significant variations across the three subject groups. To determine the significance, we employed the Kruskal-Wallis H-test. Additionally, pairwise comparisons between the different subject groups were performed using Dunns' tests described in the previous section. These analyses allowed us to assess and compare the patterns of variation across the different distance measures and profile types.

2.2.3 Identification of diagnostic markers and differentially abundant taxa

We employed two parallel approaches to identify taxonomic markers with different groups' diagnostic features. In the first approach, we utilized Random Forest to de-

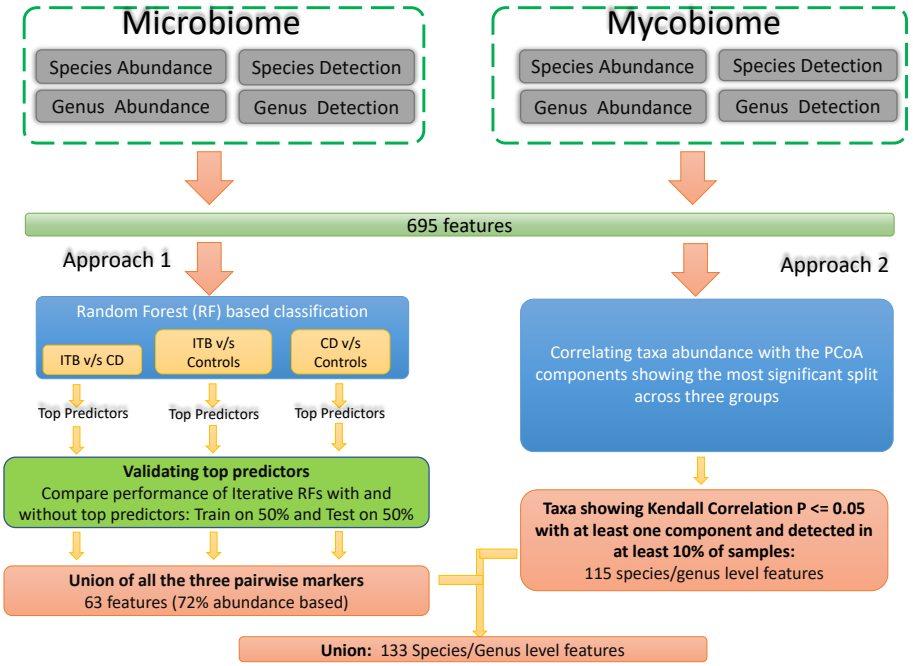

Figure 2.4: Diagnostic marker identification

termine the top diagnostic predictors for discriminating between the three group pairs (Controls vs. ITB; Controls vs. CD; CD vs. ITB) [Figure 2.4]. The approach's details are provided below.

First, we compiled a comprehensive matrix that included all the microbiome and mycobiome profiles, along with the abundance and detection values of all species and genus-level taxa at both levels. This matrix resulted in 53 samples and 695 microbiome features. Next, we constructed a single 'Random Forest classifier' for a specific pair of groups. The classifier aimed to classify the samples into two groups based on the 695 features. This classification process was performed using the 'randomForest' function from the 'randomForest' package v4.6.14.

After constructing the Random Forest models, we ranked the 695 features based on their feature importance scores, representing the mean GINI decrease across 500 iterations. Using these ranks, we generated multiple Random Forest models considering various top features (e.g., top 10, 20, 30, 40, up to 250). To identify the most discriminatory and reproducible features for classifying the two groups in each pairwise scenario, we selected the number of features for which the corresponding Random Forest models generated the highest Area Under the Curve (AUC) values. We ensured that

biases in certain samples did not affect this set of diagnostic features. For this purpose, we performed 50 iterations, where each iteration involved selecting 50% of the samples for training while the remaining 50% was reserved for testing. For each iteration, two variants of Random Forest models were created: one using the top predictors identified earlier for each pair of groups and the other using all other features apart from the top predictors. We then compared the AUCs of classification obtained for these two variants on the 50% test sets across the 50 iterations using Wilcoxon tests. This comparison allowed us to confirm if the AUCs obtained for the former variants (top predictors) were significantly higher than those obtained using the latter (all other predictors) for all three pairwise scenarios. The validation process ensured the robustness and reliability of the selected diagnostic features.

In the second approach, we focused on identifying taxonomic features (species and genus abundances at both microbiome and mycobiome levels that exhibited significant correlations with the Principal Coordinates displaying the most significant axis of splits ($|\text{Kendall tau}| \geq 0.20$ and p-value ≤ 0.05) [Figure 2.4]. Combining the results of both approaches, we obtained a comprehensive set of 133 features. These features comprised the abundance levels of specific microbiome and mycobiome taxa at the species and genus levels, identified using the two approaches.

2.2.4 Understanding the mutual relationships with Biomarkers with the network-based approach

Next, we examined the community-level mutual co-abundance relationships among the aforementioned diagnostic markers. We aimed to explore co-abundant modules of diagnostic markers specific to each subject group (Controls, ITB, and CD). We employed the permutation-renormalization bootstrap (ReBoot) approach, implemented through the 'ccrepe' module version 1.34.0 of the R programming interface [48]. We used the Kendall-tau correlation score for assessing the co-abundance relationships between markers.

In our analysis, we identified co-abundance relationships (Kendall-tau > 0) with a false discovery rate (FDR) of ≤ 0.1. These relationships were then represented as a marker versus marker adjacency matrix. To further visualize these relationships, we constructed a network representation. In this network, each marker with co-abundance relationships with at least one other marker was depicted as a node. An edge also connects any two nodes with a mutual co-abundant relationship. We employed the 'igraph' package v1.35.0 in R to create this network representation, specifically utilizing the 'graph_from_adjacency_matrix' function. This approach gave us insights into the interplay and co-abundant patterns among the diagnostic markers specific to each

subject group.

2.2.5 Validating reproducibility of marker relationships in global meta networks using multiple publicly available microbiome and mycobiome data sets

We implemented a meta-network as a graph representing the co-abundance relationships among taxonomic features inferred from conducting meta-analysis across multiple studies. In this context, the taxonomic features served as nodes in the network. The meta-analysis approach was adopted to determine the connections between the nodes (features).To achieve this, we initially computed the abundance association between each pair of taxa across all the studies. These abundance associations were subsequently examined using the meta-analytic Random Effects Model to derive p-values and the directionality of association across the studies. For any given taxa, the p-values with all others were corrected using the Benjamini-Hochberg method to compute the False Discovery Rates (FDRs). Only two features were connected by an edge if both exhibited a significant positive association (Summarized Estimate $> 0; \text{FDR} \leq 0.1$) in a meta-analysis performed across multiple studies. Furthermore, this summarized association had to be reproduced with the same directionality across at least 70% of the study cohorts. To conduct the meta-analysis using a random-effects model, we employed the 'metafor' package v2.0 and the 'robumeta' package v3.8 in R. The resulting meta-network provided valuable insights into the co-abundance patterns among taxonomic features identified from the meta-analysis across multiple studies.

We conducted separate meta-analyses to explore three types of associations: intra-microbiome (focused on gut microbiome profiles only), intra-mycobiome (focused on gut mycobiome profiles only), and mycobiome-microbiome co-abundance associations (focused on paired. profiles with both gut microbiome and mycobiome data). Our study encompassed a total of 14 studies, comprising 5437 gut microbiome profiles and 892 gut mycobiome profiles [Table 2.1]. For the microbiome profiles, we primarily included data sets from studies that exclusively focused on the Indian population[16–23]. These studies contributed a total of 1617 gut microbiome profiles from India. Additionally, we incorporated 3559 gut microbiome profiles from apparently healthy individuals in a population-level cohort from neighboring China[30]. Moreover, one of the data sets also contained matched gut mycobiome profiles from both an Asian (specifically Japanese) population and India (total N=97)[23].

We further included two more data sets from the European Union that focused on investigating IBD and IBS (irritable bowel syndrome), which also encompassed both gut microbiome and mycobiome profiles[24–26]. We utilized these paired gut micro-

biome and mycobiome profiles for meta-analysis investigating microbiome-mycobiome associations. This set of mycobiome profiles was supplemented with three additional data sets from the Human Microbiome Project v2, originating from India and China, to investigate intra-mycobiome associations[27, 31]. Network edges identified within our current study cohort (as described in the previous step) also observed in the three meta-analyses performed in this step were identified as the core set of reproducible associations. These core associations represented robust co-abundance relationships consistently observed across different data sets and populations.

2.2.6 Identification of an Indian core microbiome and computation of the core gut microbiome score (CGMS)

We collected a total of 1617 gut microbiomes from seven studies, with a specific focus on Indian sub-populations (as previously described). To identify the core gut taxa within the Indian population, we utilized two properties: Prevalence and Centrality. For the Prevalence-based approach, we identified prevalent taxa as those present in at least 70% of the samples in at least four of the seven data sets. To identify central taxa within the Indian population, we constructed a meta-network using only the seven India-based studies, similar to the approach described earlier. Within this network, taxa ranked in the 70th percentile in terms of their degree of centrality were identified as the central taxa. Finally, the core gut taxa within the Indian population were determined as the taxa that appeared in both the prevalent and central taxa lists. These taxa exhibited high prevalence and centrality within the Indian sub-population, making them important and prominent contributors to the core gut microbiome in this population.

After identifying the core gut microbiome, the core gut taxa were sorted in descending order based on their meta-network degree centrality. Regardless of whether it belonged to the previously mentioned list of 1617 microbes, all core gut taxa were identified for any given gut microbiome data set. These taxa were then ranked across samples and then rank-scaled from 0 to 1 using the below formula:

The rank-scaled abundance of a given taxa 'j' in sample 'i'=
((rank (j in i) − min(rank(j across all i)))/
((max(rank(j across all i)) − min(rank(j across all i)))

The core gut microbiome score for a data set was then determined as below:
For a given microbiome k:

$$\textbf{CGMS} = \sum_{\textbf{all detected core taxa } m} \textbf{Rank-Scaled Abundance of taxa } m \text{ in } k$$

* Percentile rank of 'm' in degree centrality in the meta-network identified above.

After identifying the Indian core microbiome species, we validated the core gut microbiome score to determine the disease-associated core microbiome depletion by utilizing this core species in seven data sets with a total microbiome size of 3030[Table 2.1].

2.2.7 Resiliency analysis, including longitudinal studies with multi-timepoint sampling

Our next step was to check the depletion pattern of the microbiome not at a single time point of study but of studies at multiple time points. This multi-time point study aimed to review the resiliency of the microbiome in samples across different observational time points. The criteria to check the resiliency was verifying the microbiome's detection and abundance variation across multiple time points and the effect of the core group of gut microbiomes on this diversity throughout these time points. To do so, we first collected clinical information from 12 gut microbial studies across different parts of the world. The combined studies resulted in 5057-time point samples from 1783 subjects with various gut-related infections and also including healthy controls. [Table 2.2].

2.2.8 Data preprocessing and Metadata verification for longitudinal studies

We downloaded the patient-related metadata from the study and the relative genomic sequence data from the European Nucleotide Archive[Figure 2.5]. The 16S and shotgun metagenomic sequences were taxonomically classified, and the abundance profile matrices were generated. We implemented the SPINGO [45]. microbial amplicon classifier for the 16S amplicon sequence reads, and the shotgun metagenomic reads were classified using MetaPhlAn 4.0 [49] taxonomic profiler. The classified species abundance profile matrix was then compared with the metadata collected from the respective studies to find only longitudinal samples and their species abundances. For each analysis, a matrix with each row containing the longitudinal sample at a time 'T' and its follow-up sample at a time 'T+1' was prepared and was used as a reference for calculating the species abundance variation among the specimens and their follow-up. This variation of the abundance displayed the extent of resiliency in the species across multiple time points in a subject. Intuitively the variation in the abundance of resilient species should be less among the sample and the follow-up.

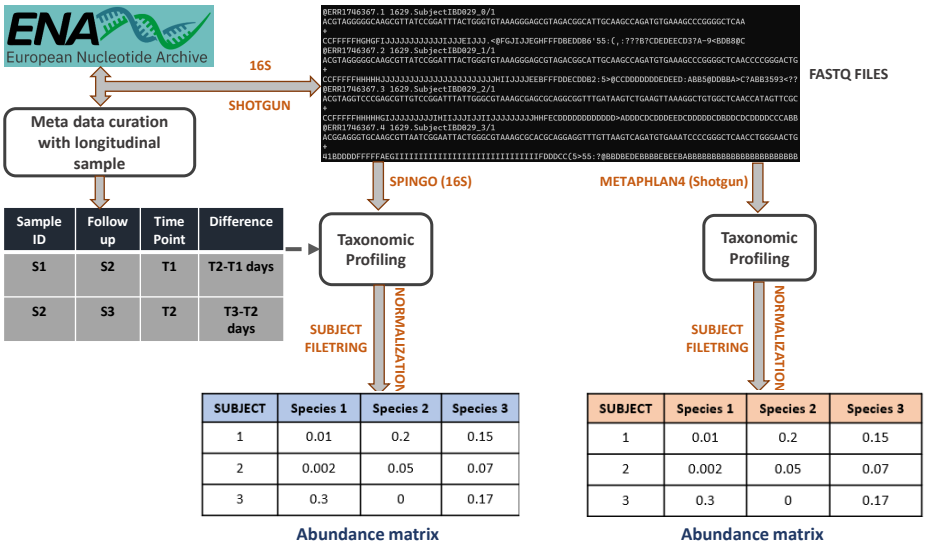

Figure 2.5: Data preparation pipeline for longitudinal samples

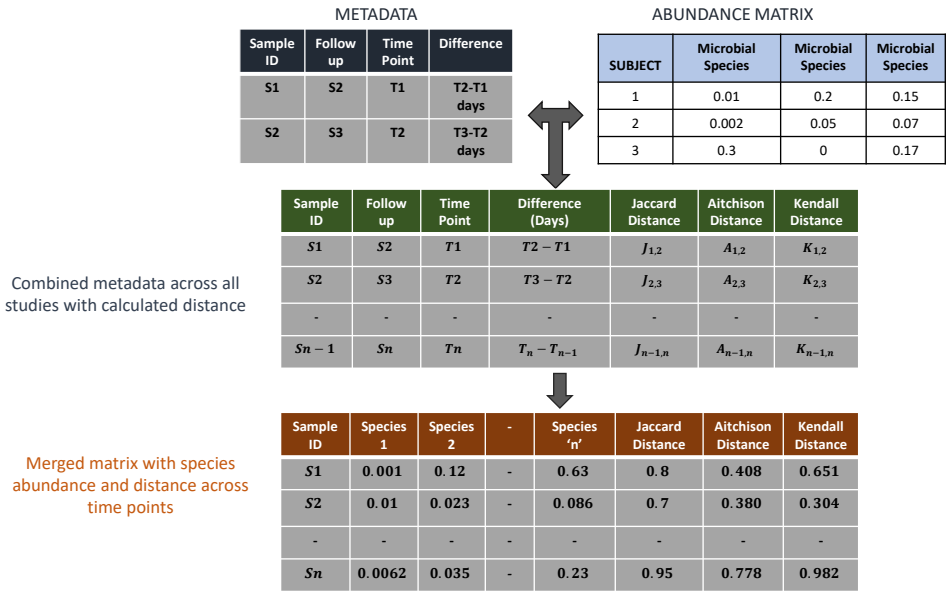

Figure 2.6: Species abundance to distance matrix generation

After the data preprocessing step, we calculated the distance among the longitudi-

nal samples[Figure 2.6]. The species abundance matrix was initially transformed using the formula 'x-min(x)' where 'x' is the species abundance and 'min(x)' is the minimum of the species abundance in a given sample to cope with the zero and near zero values. Then the abundance matrices were normalized using two methods; in the first method, we did the row sum scaling to get each species' relative abundance, and in the second, we implemented the CLR transformation using the 'clr' function of the compositions package in R v2.0.4. As mentioned in module-1, three distance calculation methods (Kendall, Weighted-Jaccard, and Aitchison) were adopted in our resiliency study to assess the detection and abundance variation. The distance across the time point was computed below:

Distance $(S_T, S_{T+1}) = \{J_{T,T+1}, A_{T,T+1}, K_{T,T+1}\}$

where S_T is a sample at time point T, S_{T+1} is the follow-up sample of S_T at time point *T+1*. $\{J_{T,T+1}, A_{T,T+1}, K_{T,T+1}\}$ are the Jaccard distance, Aitchison distance, and Kendall distance, respectively.

2.2.10 Estimation of the effect of the core groups' abundance on the diversity of the microbiome community

The calculated distance columns and the study details were added to the longitudinal sample matrix. Each study's normalized abundance matrices were combined with the distance data available with three distance measurements, and the combined data were then used in the meta-analysis step. In the meta-analysis, we used the 'rlm' function from the 'MASS' package in R for robust linear regression analysis. This regression analysis of species abundance with the "Kendall", "Jaccard," and "Aitchison" distance among the time point samples in a data set produced the regression coefficients for effect size calculation. The 'metafor' package in R consisting of the 'escalc' function was used to calculate the study-wise effect size. Finally, the 'rma' function from the same 'metafor' package was used to generate the overall effect size, including all of the longitudinal studies. A matrix consisting of the overall summarised estimates, associated p values, FDR corrected p values, confidence intervals, and overall directionalities of the species effects was generated. The directionality with high magnitudes produced a high effect on the stability of a particular taxon across different time points.

We calculated study-specific effect estimates for all 12 studies and the overall directionality of all the species. As mentioned in module-1, the core microbiome loss, particularly in IBD cases, was the key player in the breakage of co-abundance networks of healthy microbiomes. To identify the effects of the resiliency of the health-associated microbiomes in longitudinal samples, we have used our meta-analysis results for the 28

species identified in the Indian core gut microbiota. This set of species was then used to find the overall summarised effect estimates and directionality with respect to the microbiome diversity in our meta-analysis result. The effect and directionality of these sets of species were also verified in each of the individual data sets, where we did the robust linear regression analysis of the species abundances with three of the distance measures mentioned above for each data set.

Chapter 3

RESULTS

3.1 Results of the group-specific dysbiosis analysis in the samples from AIIMS

3.1.1 Intestinal Tuberculosis and Crohn's Disease are characterized by distinct microbiome- and mycobiome-level alpha and beta-diversity variations with respect to controls

In this particular study cohort, we collected gut microbiome profiles from 57 subjects, with 53 of these subjects having matching gut mycobiome profiles. The subjects were categorized into three groups: Controls, Crohn's Disease (CD), and Intestinal Tuberculosis (ITB). To understand the differences in the alpha diversity of the gut microbiomes, we utilized two indices: the Shannon Index, which provides a measure of both the richness and evenness of the gut microbiome and Pielou's index, which specifically quantifies the evenness of the gut microbiome. We performed this analysis at both microbiome and mycobiome levels [Figure 3.1].

Comparing the microbiome Shannon indices, we observed a distinct decrease in the CD and ITB groups when compared to the controls. The difference became even more pronounced when analyzing the Pielou Evenness indices, indicating a significant decrease in the gut microbiome evenness in the CD and ITB groups as compared to the controls. Notably, there was no significant variation in gut microbiome evenness between the CD and ITB groups. This significant decrease in evenness for the diseases was consistently replicated across all three taxonomic levels (ASV, Species, and Genus)[Figure 3.1]. However, at the mycobiome level, the observed pattern was no-

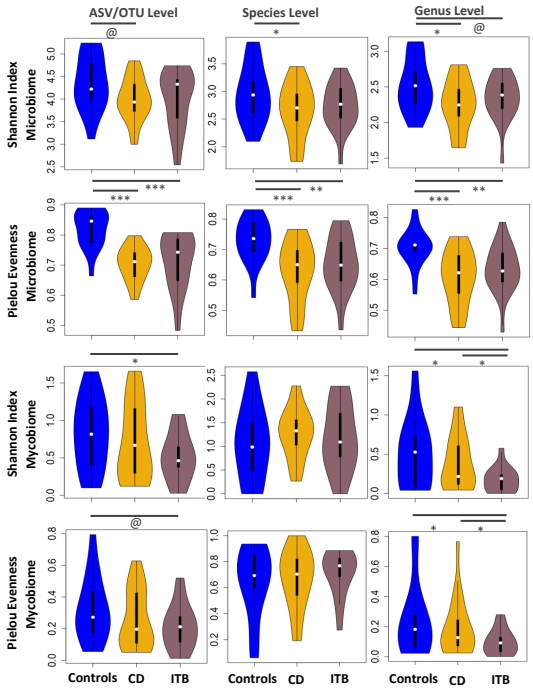

Figure 3.1: Alpha diversity result of 03 groups(ITB, CD, and Control) showing a reduction in Shannon index and Pielou's evenness in ITB and CD at microbiome level. CD showed distinct increment indices at the mycobiome level, whereas ITB had decreased indices.

tably different. Compared to Intestinal Tuberculosis (ITB), the Controls exhibited significantly higher levels of the Shannon Index (both at the OTU and Genus levels). This finding suggests that ITB subjects have a distinctive reduction in the fungal (or "mycobial") community as compared to the Controls. On the other hand, the Crohn's Disease (CD) group displayed a distinct mycobiome diversity pattern resembling that of the Controls. In fact, at the species level, the mycobiome diversity in the CD group was non-significantly higher than that in the Controls, and it was notably higher than in the ITB group (significant at the Genus level). This observation indicates that although ITB and CD, two diseases with similar clinical symptom presentations, exhibit similar gut microbiome alterations, they display distinct changes at the mycobiome level. This distinction is particularly important since an increased abundance of fungal lineages has previously been associated with Inflammatory Bowel Diseases like Crohn's [50]. The unique decrease in mycobiome diversity in this cohort of ITB patients could represent a distinctive pattern of dysbiosis, which might serve as a differentiating factor between the two diseases.

A series of Principal Coordinate Analyses (PCoAs) was conducted to explore the variation patterns in the overall compositions of the gut microbiome and mycobiome across the three distinct groups of individuals. These PCoAs were performed separately

Table 3.1: Significant PCoA axes discriminating (ITB, CD, and Controls): microbial species

Distance Measure	Kruskal-Wallis 'p' value		
	PCo1	PCo2	PCo3
Kendall	**0.00002**	0.4	0.9
Jaccard	**0.00002**	0.9	**0.09**
Aitchison	**0.00009**	0.1	**0.02**

Table 3.2: Pairwise group discriminating PCoA axes: microbial species(Dunn's test)

Distance Measure		ITB-CD	ITB-Controls	CD - Controls
Kendall-PCo1	Z_value	0.3	4.2	3.9
	P_value	0.5	**0.00004**	**0.0001**
Kendall-PCo2	Z_value	-0.6	0.8	1.4
	P_value	0.5	0.3	0.1
Kendall-PCo3	Z_value	-0.1	0.3	0.4
	P_value	0.5	0.4	0.4
Jaccard-PCo1	Z_value	0.4	4.2	3.9
	P_value	0.4	**0.00003**	**0.0002**
Jaccard-PCo2	Z_value	0.2	-0.2	-0.4
	P_value	0.4	0.4	0.4
Jaccard-PCo3	Z_value	0.2	-1.8	-2
	P_value	0.4	0.1	**0.03**
Aitchison-PCo1	Z_value	0.4	4	3.6
	P_value	0.5	**0.0001**	**0.0006**
Aitchison-PCo2	Z_value	0.4	1.9	1.5
	P_value	0.5	0.03	0.06
Aitchison-PCo3	Z_value	-0.1	2.4	2.5
	P_value	0.5	**0.01**	**0.01**

Table 3.3: Significant PCoA axes discriminating (ITB, CD, and Controls): microbial genus)

Distance Measure	Kruskal-Wallis p_value		
	PCo1	PCo2	PCo3
Kendall	**0.00002**	**0.2**	**0.093**
Jaccard	**0.00006**	0.9	**0.0007**
Aitchison	**0.00002**	0.5	**0.0435**

Table 3.4: Pairwise group discriminating PCoA axes: microbial genus(Dunn's test)

Distance Measure		ITB-CD	ITB-Controls	CD-Controls
Kendall-PCo1	Z_value	-0.65	-4.4	-3.73
	P_value	0.31	**0.00002**	**0.0003**
Kendall-PCo2	Z_value	0.5	-1.2	-1.73
	P_value	0.31	0.2	**0.04**
Kendall-PCo3	Z_value	1.58	-0.6	-2.07
	P_value	0.17	0.3	**0.03**
Jaccard-PCo1	Z_value	0.72	4.2	3.48
	P_value	0.35	**0.00004**	**0.0004**
Jaccard-PCo2	Z_value	-0.14	0.3	0.5
	P_value	0.44	0.4	0.3
Jaccard-PCo3	Z_value	1.03	-2.7	-3.7
	P_value	0.35	**0.005**	**0.0003**
Aitchison-PCo1	Z_value	0.45	4.2	3.8
	P_value	0.41	**0.00003**	**0.0002**
Aitchison-PCo2	Z_value	-0.72	-1.14	-0.4
	P_value	0.41	0.13	0.3
Aitchison-PCo3	Z_value	0.22	2.29	2.1
	P_value	0.41	**0.02**	**0.03**

Table 3.5: Significant PCoA axes discriminating (ITB, CD, and Controls): mycobial species

Distance Measure	Kruskal-Wallis p_value		
	PCo1	PCo2	PCo3
Kendall	0.7	**0.01**	0.2
Jaccard	0.5	0.2	0.2
Aitchison	0.2	0.8	0.9

Table 3.6: Pairwise group discriminating PCoA axes: mycobial species (Dunn's test)

Distance Measure		ITB-CD	ITB-Controls	CD - Controls
Kendall-PCo1	Z_value	0.9	0.7	-0.1
	P_value	0.2	**0.2**	**0.5**
Kendall-PCo2	Z_value	2.3	2.8	0.6
	P_value	0.03	0.01	**0.5**
Kendall-PCo3	Z_value	1.7	1.2	-0.5
	P_value	0.1	0.2	**0.5**
Jaccard-PCo1	Z_value	1.1	1.1	0.1
	P_value	0.1	**0.1**	**0.5**
Jaccard-PCo2	Z_value	-1.8	-1.6	0.1
	P_value	0.06	0.09	0.5
Jaccard-PCo3	Z_value	-1.8	-1.5	0.2
	P_value	0.06	**0.09**	**0.5**
Aitchison-PCo1	Z_value	-1.7	-1.3	0.3
	P_value	0.1	**0.3**	**0.5**
Aitchison-PCo2	Z_value	0.5	0.4	-0.09
	P_value	0.3	0.5	0.5
Aitchison-PCo3	Z_value	-0.4	-0.07	0.4
	P_value	0.3	**0.5**	**0.5**

Table 3.7: Significant PCoA axes discriminating (ITB, CD, and Controls): mycobial genus

Distance Measure	Kruskal-Wallis p_value		
	PCo1	PCo2	PCo3
Kendall	**0.04**	0.5	0.13
Jaccard	**0.05**	0.64	0.12
Aitchison	0.11	0.53	0.62

Table 3.8: Pairwise group discriminating PCoA axes: mycobial genus (Dunn's test)

Distance Measure		ITB-CD	ITB-Controls	CD - Controls
Kendall-PCo1	Z_value	-2.05	-2.3	-0.4
	P_value	**0.06**	**0.03**	**0.4**
Kendall-PCo2	Z_value	-1.2	-0.7	0.4
	P_value	0.2	0.2	**0.4**
Kendall-PCo3	Z_value	-0.3	1.6	1.9
	P_value	0.4	0.1	**0.08**
Jaccard-PCo1	Z_value	2.3	1.9	-0.3
	P_value	**0.04**	**0.1**	**0.4**
Jaccard-PCo2	Z_value	0.9	0.4	-0.5
	P_value	0.3	0.3	0.4
Jaccard-PCo3	Z_value	0.2	-1.7	-1.9
	P_value	0.4	**0.1**	**0.1**
Aitchison-PCo1	Z_value	-2	-1.6	0.3
	P_value	0.07	**0.2**	**0.4**
Aitchison-PCo2	Z_value	0.5	1.1	0.7
	P_value	0.3	0.2	0.4
Aitchison-PCo3	Z_value	-0.8	-0.9	-0.1
	P_value	0.3	**0.2**	**0.4**

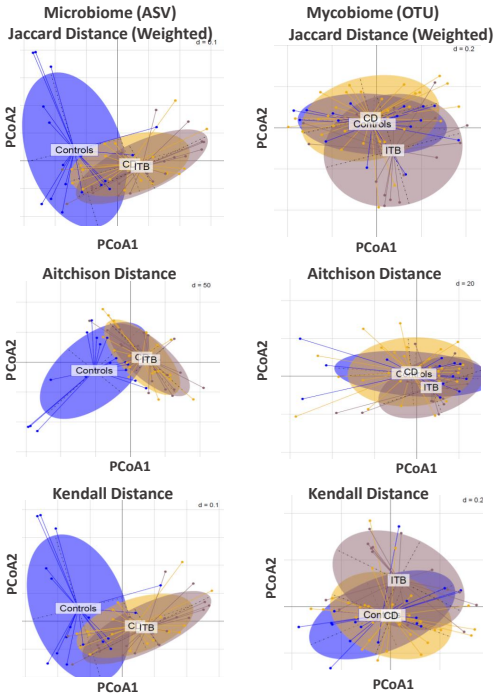

Figure 3.2: Beta diversity result of 03 groups at microbial ASV and mycobial OTUs

for the composition profiles obtained at each of the three taxonomic levels (OTU/ASV, Species, and Genus) using three different distance matrices (Weighted Jaccard, Aitchison, and Kendall) [Figure 3.2,3.3,3.4]. The association data obtained from PCoAs for the microbiomes of each group are available in [Table 3.1,3.2] at the species level and [Table 3.3,3.4] at the genus level. For mycobiome profiles, the corresponding data can be found in [Table 3.5,3.6] for species and [Table 3.3, 3.8] for the genus. It was observed that the gut microbiome composition of individuals with Crohn's Disease (CD) and Intestinal Tuberculosis (ITB) exhibited similarities and was significantly distinct from that of the control group. However, in contrast, while the composition of the mycobiome in CD patients was similar to that of the control group, the ITB patients showed a distinct mycobiome composition.[Figure 2.3,3.3,3.4]. These patterns were consistent across all taxonomic levels and distance measures used.

This investigation's results and the alpha diversity analysis findings indicated variations in the overall disease-associated alterations between CD and ITB. Specific patterns were observed at both the microbiome and mycobiome levels, highlighting the uniqueness of each disease's gut microbial composition.

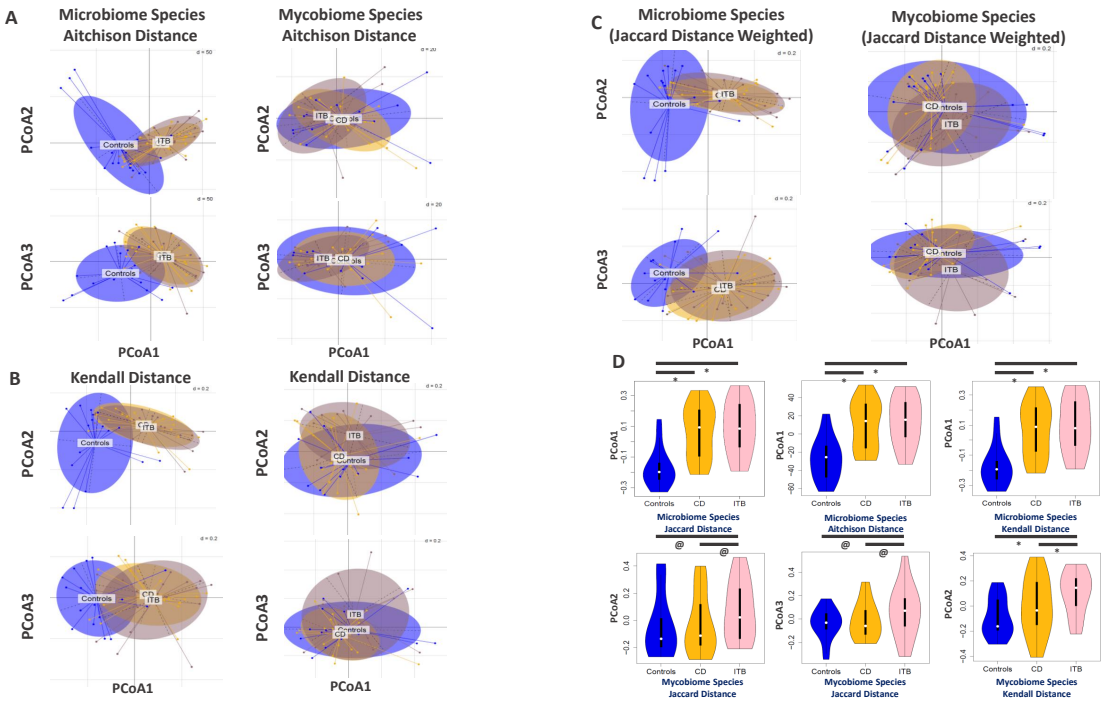

Figure 3.3: Beta diversity result of 03 groups at the microbial and mycobial species level. A, B, and C explain the overall diversity and associated significant PCoA axis of the split. D represents the significant axes describing the pair-wise diversity resulting from Dunn's test

3.1.2 Distinct taxonomic markers enable pairwise discrimination within the two different diseases and between the two disease groups and controls.

In the subsequent phase of our investigation, we aimed to determine the extent to which the gut microbiome of the three groups of individuals could differentiate between the two diseases and distinguish the diseases from the control group. Additionally, we tried to identify specific taxa (or modules or guilds of particular taxa) that could play a crucial role in distinguishing between the three groups. To achieve this objective, we adopted a combined approach that involved two different methods. Firstly, we utilized Random Forest-based pairwise group classification to discern between the groups. This method helped us identify key predictors that contributed to classifying subjects into their respective groups. Secondly, we sought to identify taxonomic groups associated with Principal Coordinate Analyses (PCoAs) that exhibited the most significant axis

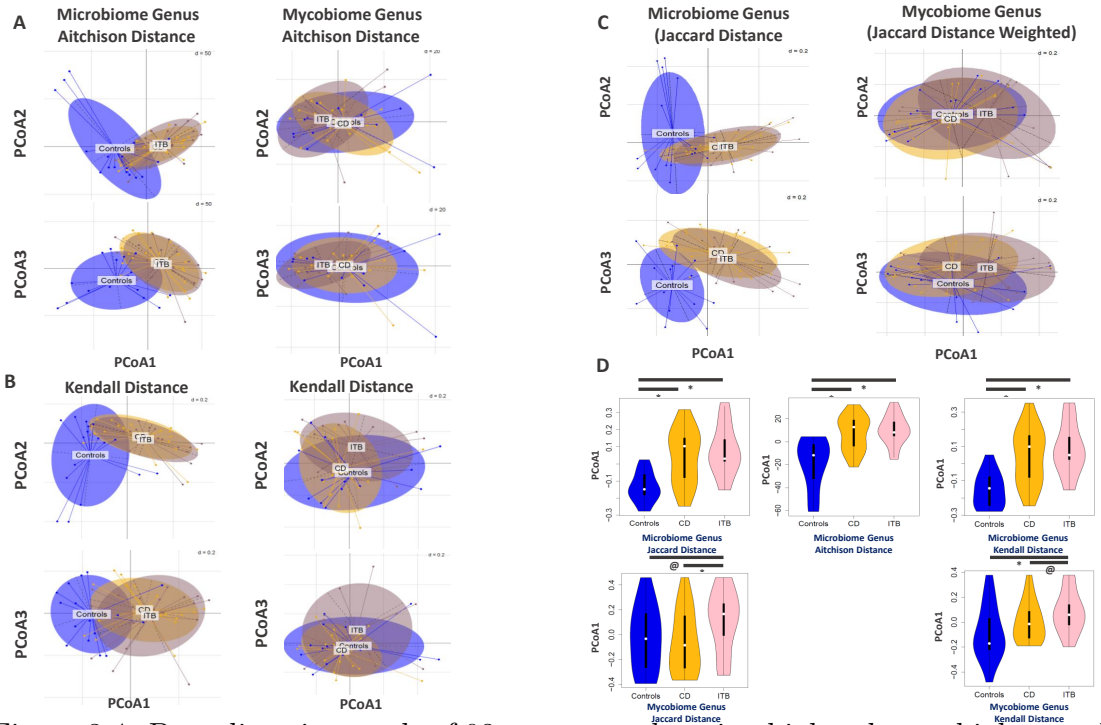

Figure 3.4: Beta diversity result of 03 groups at the microbial and mycobial genus level. A, B, and C explain the overall diversity and associated significant PCoA axis of the split. D represents the significant axes describing the pair-wise diversity resulting from Dunn's test

of split [Figure 3.3D, 3.4D, and 2.4], as well as Table [3.1,3.2,3.3,3.4,3.5,3.6,3.7,3.8].

Through the execution of multiple Random Forest models, each considering a varying number of top features (ranging from 10 to 250), we successfully identified a set of 10, 40, and 30 most discriminatory taxonomic features. These sets achieved the highest diagnostic accuracy for pairwise discrimination between Controls vs. Intestinal Tuberculosis (ITB) (top AUC: 96.9%), Controls vs. Crohn's Disease (CD) (top AUC: 95%), and CD vs. ITB (top AUC: 80%)[Figure 3.5A]. These findings demonstrated that the gut microbiome profile exhibited an exceptional ability to distinguish between the patients (CD/ITB) and healthy Controls with very high efficiency. Moreover, it could also effectively differentiate between the two diseased groups (CD and ITB) with a reasonable level of accuracy. This discovery highlighted the significant potential of the gut microbiome as a valuable tool for disease discrimination and its promise to provide valuable insights into the complex interactions between the gut microbiota and different disease conditions.

For each pair of groups, we further validated these markers using iterative bootstrapped Random Forest models (with two-fold cross-validation; See Methods), where in each iteration, we considered two variants of Random Forest models, one using the corresponding top predictors as identified in the previous step. The other uses all

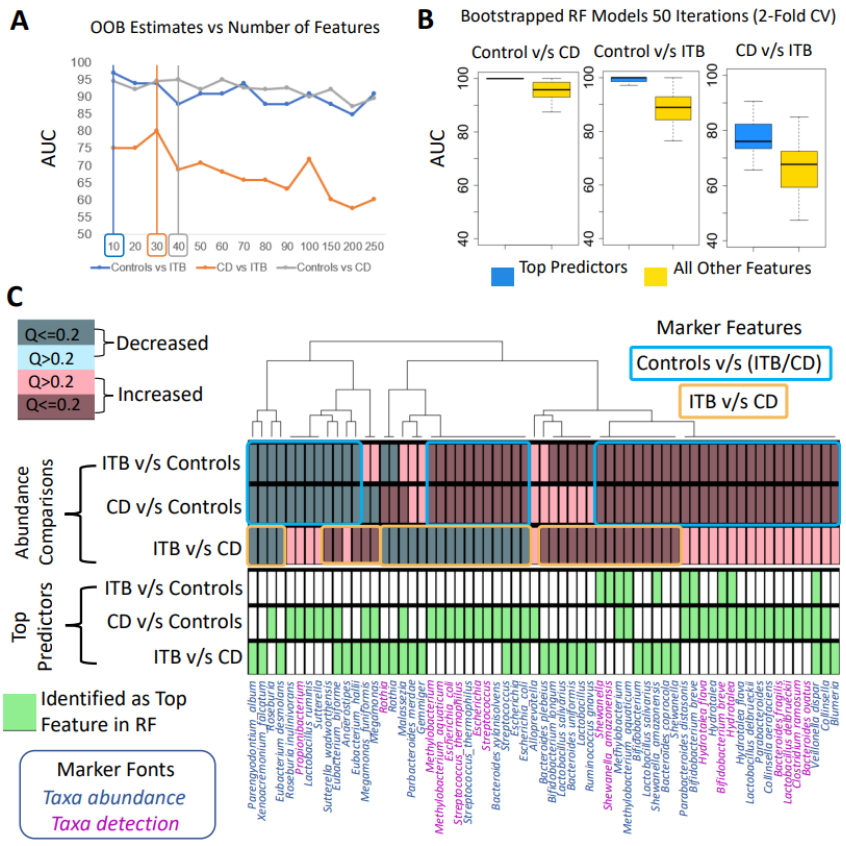

Figure 3.5: Distinct diagnostic markers using Random Forest model

taxonomic features besides the top predictors [Figure 3.5B]. For each pairwise group scenario, model variants considering the top features were observed to have significantly higher accuracy than those trained on all other features. This indicated that the identified set of top diagnostic features was robust for this cohort, not reflections of biases associated with specific "outlier" samples with specific microbial composition.

We investigated these sets of the most diagnostic features and analyzed their abundance variations across the different groups (as shown in [Figure 3.5C]. Based on their discriminating ability across groups, these features could be categorized into two main groups: features capable of distinguishing between Controls and the two diseases (CD and ITB) and features that could distinguish between CD and ITB [Figure 3.5C]. In the first category (Controls vs. CD/ITB), we observed a decrease in the detection and/or abundance of certain fungal species, including *Parengyodontium album*, *Xenoacremonium falcatum*, and bacterial lineages like *Roseburia* (species *Roseburia inulinivorans*), various *Eubacterium* species (*Eubacterium desmolans*, *Eubacterium biforme*, *Eubacterium hallii*), *Lactobacillus ruminis*, the genus *Anaerostipes*, and other bacterial taxa such as *Propionibacterium* and *Sutterella* (species *Sutterella*

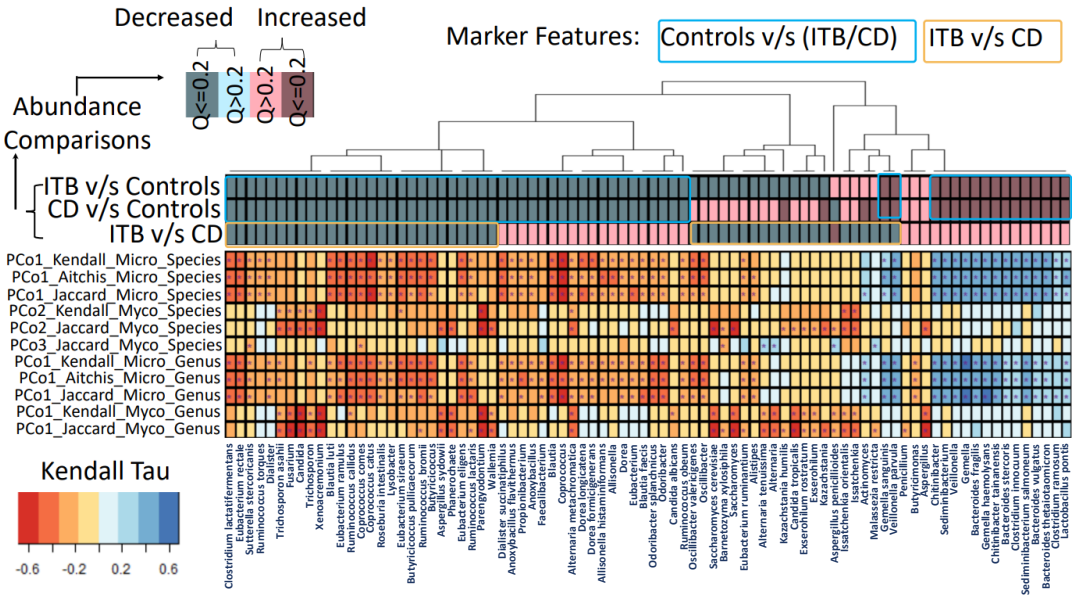

Figure 3.6: Distinct diagnostic markers using Principal Co-ordinate analysis

wadsworthensis) in both the diseases (Mann-Whitney Tests: Controls vs. CD; Controls vs. ITB; FDR $\leq$ 0.1) [Figure 3.5C]. On the other hand, we observed an increase in the detection and/or abundance of certain gut lineages, including *Methylobacterium* (*Methylobacterium aquaticum*), *Hydrotalea* (*Hydrotalea flava*), *Escherichia* (*Escherichia coli*), *Streptococcus (Streptococcus thermophilus), Shewanella (Shewanella amazonensis), Bifidobacterium (Bifidobacterium breve), Parabacteroides (Parabacteroides distasonis), Collinsella (Collinsella aerofaciens)*, various *Bacteroides* species (*Bacteroides coprocola, Bacteroides xylanisolvens, Bacteroides fragilis, Bacteroides ovatus*), multiple *Lactobacilli (Lactobacillus salivarius, Lactobacillus delbrueckii), Clostridium ramosum, Veillonella dispar*, and the fungal taxa *Blumeria* in both the diseases (Mann-Whitney Tests: Controls vs. CD; Controls vs. ITB; FDR $\leq$ 0.1) [Figure 3.5C).

The second category (CD v/s ITB) consisted of taxa *Sutterella wadsworthensis, Eubacterium biforme, Anaerostipes, Eubacterium hallii, Megamonas (Megamonas funiformis), Bacteroides plebeius, Shewanella (Shewanella amazonensis), Methylobacterium (Methylobacterium aquaticum), Bifidobacterium (Bifidobacterium longum), Lactobacillus (Lactobacillus salivarius), Bacteroides uniformis, Bacteroides coprocola, and Ruminococcus gnavus*, whose abundance and/or detection was observed to be increased

in ITB versus CD(Mann-Whitney Tests; FDR $\leq$ 0.1) and fungal taxa *Parengyodontium album, Xenoacremonium falcatum, Malassezia*, bacterial lineages *Escherichia (Escherichia coli), Streptococcus (Streptococcus thermophilus), Roseburia, Gemmiger, Eubacterium desmolans, Rothia, Parabacteroides merdae, Methylobacterium (Methylobacterium aquaticum)* and *Bacteroides xylanisolvens* were observed to be decreased in ITB v/s CD (Mann-Whitney Tests: FDR $\leq$ 0.1) [Figure 3.5C].

While the Random Forest approach efficiently identifies diagnostic markers, it may overlook some markers that could also play a diagnostic role in specific situations. In some cases, the disease association might not be solely driven by the exact diagnostic markers but also by other related gut microbial members. This was evident from [Figure 3.5B], where some pairwise classifications showed significant diagnostic markers. In contrast, the AUCs of classifier variants derived from all other features (besides the diagnostic markers) were still relatively high (greater than 0.85, especially for the classification between Controls and ITB/CD).

To identify this second set of additional features, we utilized the cross-group significantly varying Principal Coordinate Analysis (PCoA) axes and investigated the taxa associated with these axes [Figure 3.6]. By combining both approaches, we identified a total of 133 genus/species-level features at the microbiome and mycobiome levels that exhibited distinguishing capability or differential behavior for at least one of the pairwise group classifications (Controls and ITB, Controls, and CD; CD and ITB). This comprehensive analysis allowed us to identify a more extensive set of features that could potentially contribute to disease classification and provide valuable insights into the complex interactions between the gut microbiome and disease conditions. These findings open new avenues for understanding the role of specific taxa and their potential as biomarkers in different disease scenarios.

3.1.3 Diagnostic markers can be grouped into four distinct co-abundant modules in the gut microbiome

Afterward, we explored the structure of the microbial community within these gut microbiomes and investigated whether the identified diagnostic features could be categorized into separate modules showing varying patterns of changes among the three groups of subjects. To achieve this, we constructed a co-abundance network comprising the 133 taxonomic features identified earlier, using the 'ccrepe' network inference workflow specifically designed to handle compositional data (See Methods). The co-abundance network among the diagnostic taxa was divided into four distinctive modules. Among these, two modules, namely Disease Depleted Module 1 and Disease Depleted Module 2, contained taxa that were generally decreased across both diseases [Figure 3.7,3.8].

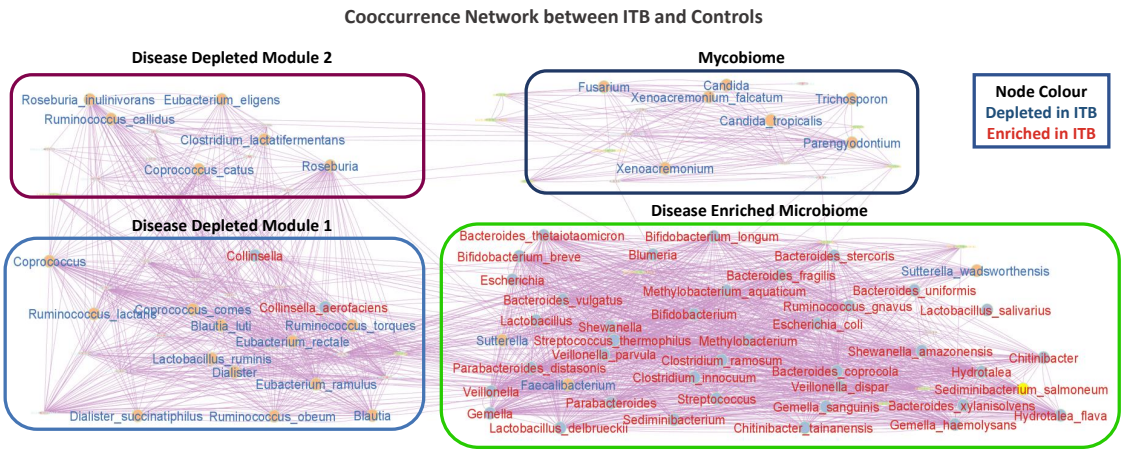

Figure 3.7: Co-occurrence network ITB v/s Controls

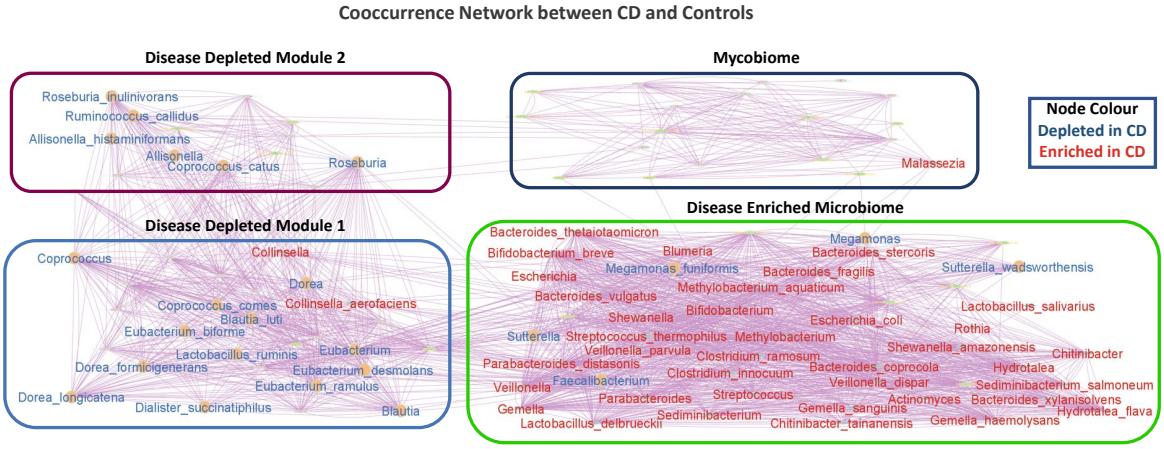

Figure 3.8: Co-occurrence network CD v/s Controls

These modules encompassed multiple taxa such as *Roseburia inulinivorans, Ruminococcus callidus, Eubacterium eligens, Coprococcus catus, Coprococcus comes, Blautia luti, Lactobacillus ruminis, Eubacterium ramulus, and Dialister succinatiphilus*, which were observed to be depleted in both CD and ITB. The overall abundance of multiple gen-

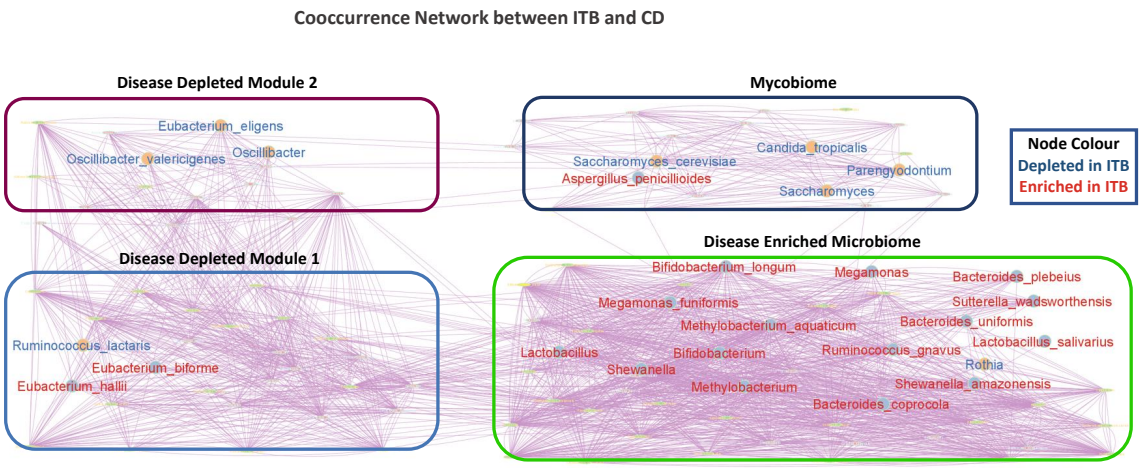

Figure 3.9: Co-occurrence network ITB v/s CD

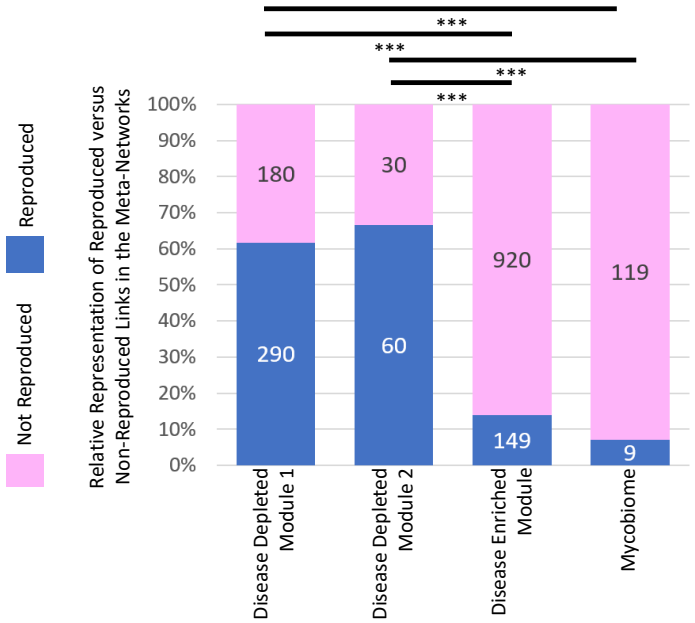

Figure 3.10: Reproducibility of co-abundance modules in global meta-networks

era like *Roseburia, Faecalibacterium, Blautia,* and *Sutterella* were also included in these modules. The third module, on the other hand, was a disease-enriched module containing multiple taxa that were observed to be enriched across both CD and ITB [Figure

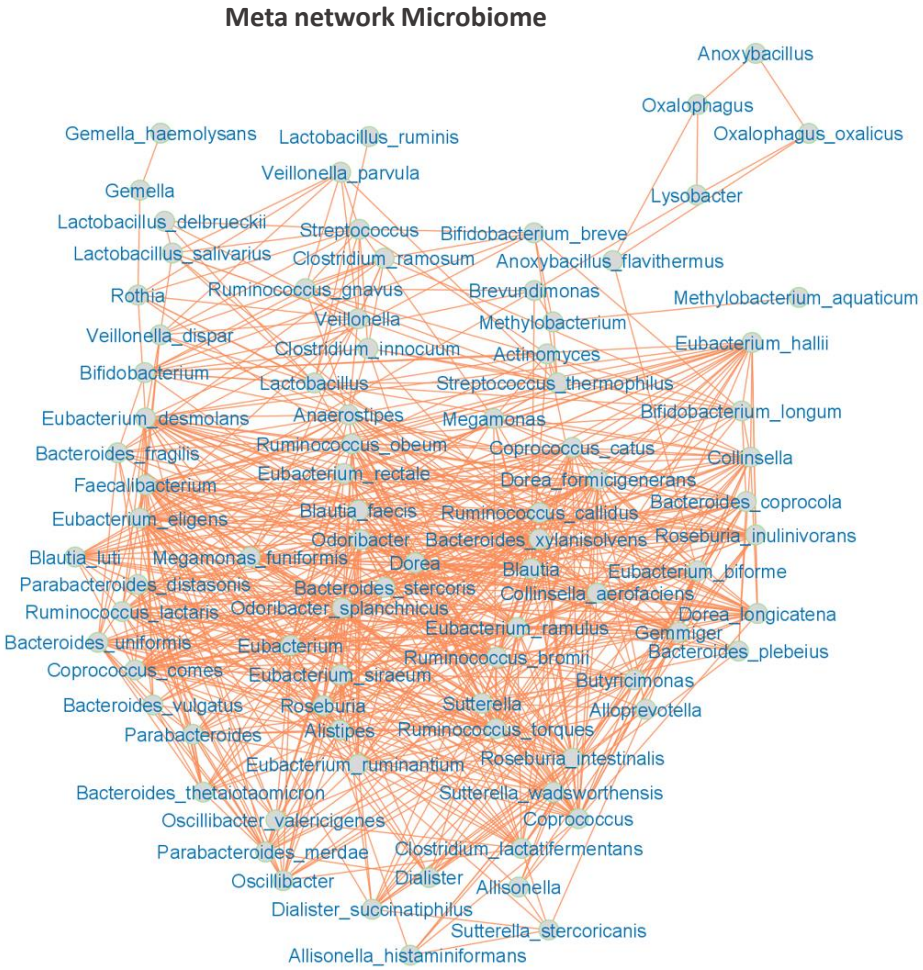

Figure 3.11: Meta network Microbiome

3.7,3.8]. Some of these taxa also showed increased abundance in ITB compared to CD [Figure 3.9]. An exception to this trend was the lineage *Megamonas* (*Megamonas funiformis*), which exhibited a decrease in CD with respect to Controls. However, a significant increase in *Megamonas* abundance in ITB compared to both Controls and CD indicated its particularly abundant presence in ITB.

The last module was dedicated to the Mycobiome and primarily consisted of fungal taxa. This module, in general, exhibited a reduced abundance in ITB when compared to both Controls and CD. However, the specific fungal taxa showing differential abundance displayed variations between Controls v/s ITB and CD v/s ITB. Most of the taxa in this module showed no significant changes between CD and Controls, except for *Malassezia*, which exhibited a significant increase in CD. Compared to CD, the only fungal species

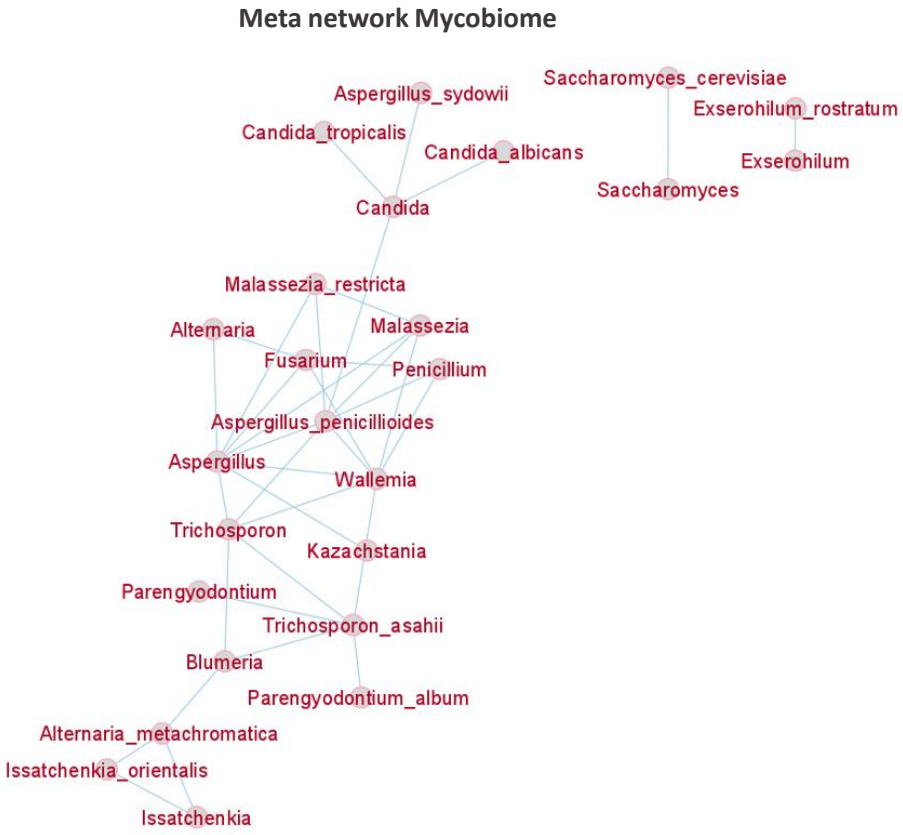

Figure 3.12: Meta network Mycobiome

observed to be enriched in ITB was *Aspergillus penicillioides*. It is worth noting that although the alterations between CD and ITB were considerably smaller than those between Controls and ITB/CD, they were still observable and restricted to specific modules within the gut co-abundance network.

We proceeded to explore the reproducibility of co-abundance associations observed within the modules identified in the global gut microbiome/mycobiomes. To assess if the disease diagnostic modules identified in our study are consistent across multiple global cohorts, which could imply their potential as universal microbiome/mycobiome targets for diagnostics or therapeutics, we constructed intra-microbiome, intra-mycobiome, and microbiome-mycobiome meta-networks. These meta-networks encompassed a total of 5,400 gut microbiome profiles and 900 gut mycobiome profiles collected from 14 different

studies worldwide (Methods module-1; 2.1). In a meta-network, each taxon served as a node, and an edge connecting two nodes if their abundances displayed a positive association based on a Random Effects Model with a false discovery rate (FDR) of ≤ 0.1 and a consistency of greater than 70% [Figure 3.11,3.12]. Subsequently, we examined the reproducibility of the intra-modular edges observed within each module identified in Figure [3.7,3.8,3.9] in the meta-networks presented in Figure [3.11,3.12].

The reproducibility of intra-modular edges varied significantly across the different modules. The two Disease Depleted Modules showed relatively high reproducibility, with more than 60% of the edges being reproduced in the meta-networks3.10. In contrast, the Disease Enriched Module and the Mycobiome displayed much lower reproducibility, with only 13.8% and 7% of the edges being consistent across the meta-networks, respectively (Fishers' Exact Test, $P \leq 0.001$). These findings suggest that the associations within the Disease Enriched Module and the Mycobiome are sporadic and observed primarily within the specific cohort of this study, potentially resulting from indirect associations caused by other changes in the gut ecology. On the other hand, the Disease Depleted Modules show greater conservation across global cohorts, indicating a specific ecological role for these Disease Depleted Taxa. Further investigations were carried out to explore this aspect in more detail.

3.1.4 Loss of the core gut microbiome is linked with the gut inflammation phenotype globally across multiple studies

Previous research has indicated the instability of the microbiome in inflammatory bowel diseases, and various studies, including the present one, have observed a reduction in diversity associated with the disease [5, 51–54]. Therefore, we aimed to determine if the two disease-depleted modules were associated with the core gut microbiome, whose loss could be linked to the inflammatory phenotype. For the first time, we attempted to identify the core gut microbiome specific to the Indian population. We gathered 1617 gut microbiomes from seven studies and utilized a combination of prevalence and meta-network degree centrality to identify a list of 46 taxa (28 species and 18 genera) that were prevalent (present in at least 70% of the samples in at least four of the seven data sets) and centrally located (degree centrality in the top 70 percentile) in the Indian gut meta-network derived from these seven studies [Figure 3.13]. Remarkably, more than 75% of the Indian core gut taxa were found in the Disease Depleted Module 1, suggesting that the loss of this core in the gut microbial community was associated with both inflammatory disease phenotypes. Subsequently, we explored the possibility of utilizing the loss of the core gut microbiome as a measure to assess the degree of gut inflammation. To achieve this, we devised a simple quantitative score termed the "core

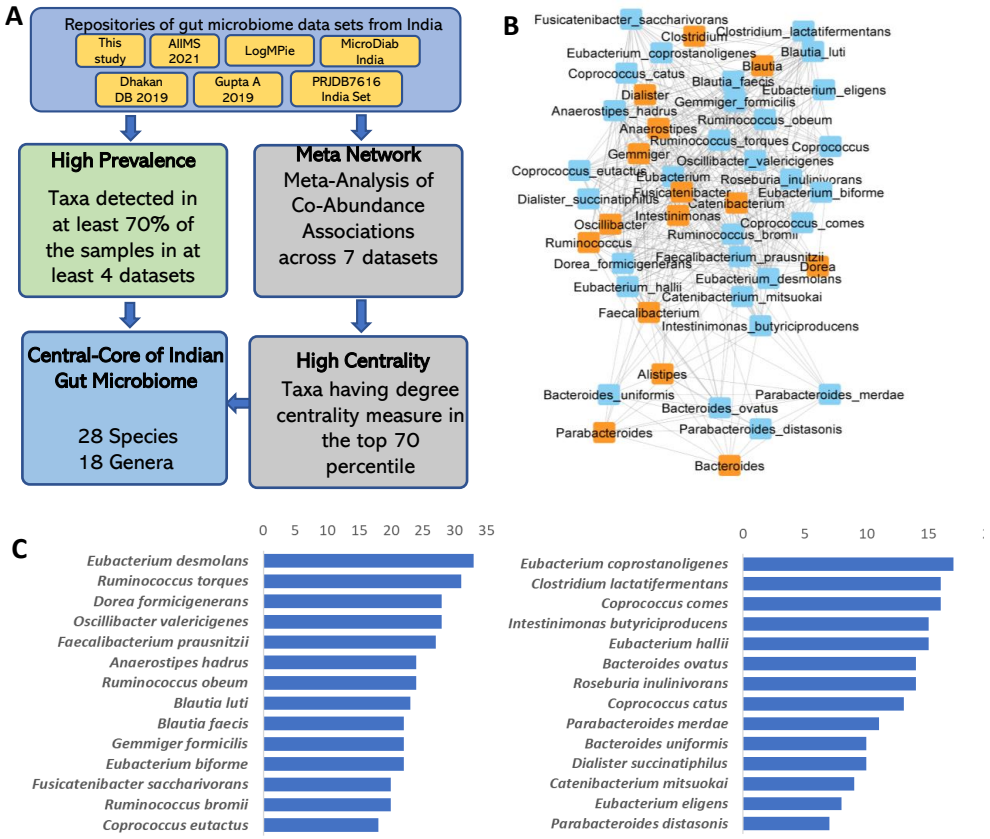

Figure 3.13: Indian core species

gut microbiome score" (CGMS). For a particular data set's microbiomes, the CGMS was calculated as the sum of the rank-scaled abundances of the core Indian gut species (Figure 14C) detected in that data set. Moreover, the abundances of each species were weighted based on its percentile in the degree of centrality within the meta-network [Figure 3.13A].

We then extended our investigation to include four study cohorts from India, Sweden, the US, and Italy, aiming to examine the variation of the CGMS across different inflammation phenotypes. In the current study, we observed that the controls had significantly higher CGMS values compared to both CD and ITB groups (with no significant difference between the two disease subtypes) (Mann-Whitney FDR ≤ 0.05). Additionally, in the previous study that explored two different inflammation phenotypes (CD/Ulcerative Colitis or UC and Acute Severe Ulcerative Colitis: ASUC), ASUC subjects exhibited higher inflammation levels than CD/UC [16]. This was also evident in

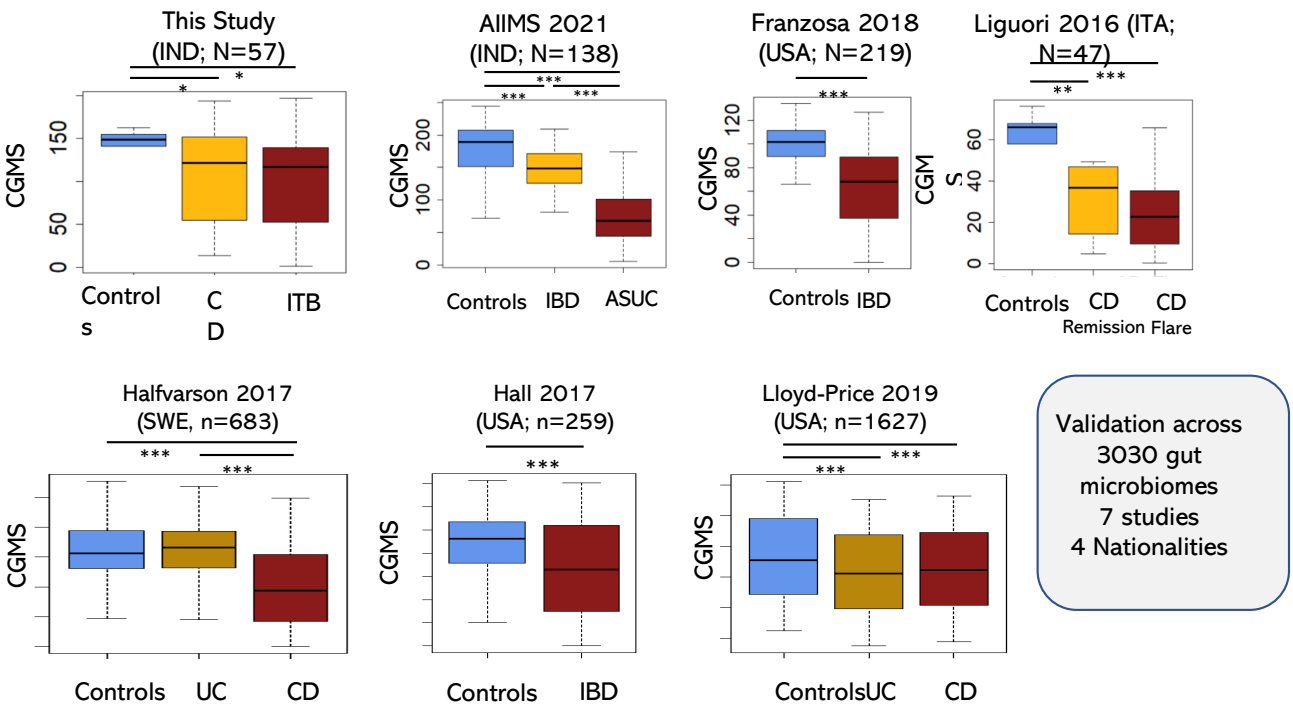

Figure 3.14: Validation of Core Gut Microbiome Score(CGMS)

the CGMS scores, as the controls had CGMS values significantly higher than both CD/UC and ASUC groups (Mann-Whitney FDR $\leq$ 0.001 for both). Furthermore, ASUC had significantly lower CGMS values than CD/UC (Mann-Whitney FDR $\leq$ 0.001), indicating a greater loss of the core gut microbiome in ASUC subjects.

We assessed whether the CGMS scores, defined based on the Indian population, could predict inflammation phenotypes in geographically distinct data sets. For this analysis, we examined the IBD cohort of the Human Microbiome Project [5]. Despite being based on the gut microbiome core of the Indian population, the CGMS scores demonstrated remarkable efficacy in distinguishing between Controls and IBD patients, with significantly lower CGMS scores observed in patients (Mann-Whitney P-value $\leq$ 0.001). We further investigated an Italian cohort [24] comprising gut microbiome profiles from Controls, IBD patients, and those in remission. Similarly, in this cohort, we found that the CGMS values of controls were significantly higher than patients and

those in remission. Moreover, individuals in remission exhibited CGMS values notably higher than the patients, indicating a partial recovery of the core microbiome. These results highlight that the abundance profiles of core gut microbial members can serve as reliable indicators to gauge the inflammation status of individuals.

We expanded our analysis to two additional cohorts from the US: one by Hall et al. (n=259) and another investigated as part of the HMP by Lloyd-Price et al. (n=1627) [5, 32]. Both cohorts consisted of gut microbiome profiles from healthy controls and patients with IBD. In both cases, significant reductions in CGMS were observed in the diseased samples compared to controls. Furthermore, we examined a Swedish cohort(n=683) [6], where a significant reduction in CGMS was also evident in CD patients. These findings underscore the potential of utilizing core gut microbiome abundance profiles as markers to measure the inflammation status of individuals across diverse populations.

3.2 Results for the resiliency in the microbiome in longitudinal studies

3.2.1 Meta-analysis across the studies resulted in a negative effect on gut microbial diversity across longitudinal studies

The combinational studies of 5057-time point samples from 1783 subjects across the 12 studies produced the extent of species' resiliency across multiple sampling time points [Table 2.2]. The combined CLR/relative abundance matrix across all the studies resulted in the union of 341 taxa. The beta diversity calculation was done using three prominent distance measures, the Kendall dissimilarity, the Jaccard distance, and the Aitchison distance, to calculate the abundance and detection level of all of the union taxa. The overall summarised effect of these 341 species with respect to each distance measure was calculated using random effect models across all the time points in all studies. This resulted in various positive and negative directionalities in the overall effects. In the case of Kendall measures, 250 species showed an overall negative directionality, while 28 species showed a negative effect with $p \leq 0.05$ and 4 with an FDR<0.1. Similarly, for the Jaccard measure, a combination of 225, 40, and 15 species was found with an overall negative effect size, with significantly negative ($p \leq 0.05$) and FDR<0.1, respectively. In the Aitchison measure, the number of species with negative effect size slightly increased; 281 species showed overall negative, 66 with $p \leq 0.05$, and 45 with FDR<0.1.

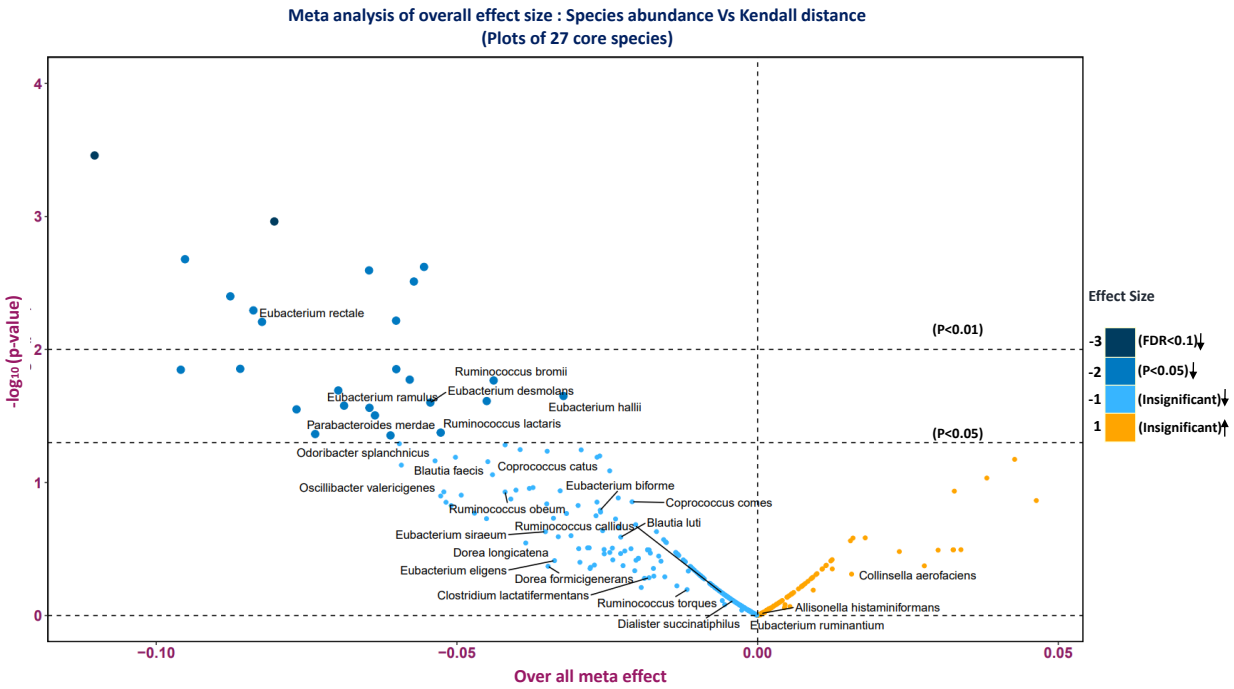

Figure 3.15: Meta effect analysis all species with respect to Kendall distance (27 species highlighted

After analyzing all the species' summarised effect sizes, we concentrated on the 28 core gut microbiomes produced in module-1 of our study. While analyzing, we found only one species, '*Lactobacillus ruminis*', which was not available in the total list of 341 taxa produced in our longitudinal data. So we then concentrated on the remaining 27 core gut species to analyze the effect estimation. These core groups' overall estimated summarised effect size and directionality were initially filtered from the meta-analysis results. 24 out of the 27 core taxa showed an overall negative effect on Kendall distance measures, whereas 8 species showed a significant negative effect with $p \leq 05$[Figure 3.15]. Regarding the Jaccard distance, we found 24 species with an overall negative effect and 11 taxa with significantly negative $p \leq 0.05$[Figure 3.16]. The effect estimation with respect to the Aitchison measure produced a negative effect for all 27 species and 6 with a significantly negative effect with $p \leq 0.05$[Figure 3.17]. The overall and study-specific meta-analysis of these core species produced the most negative summarised estimates. With the Kendall distance measure, 03 species, *Allisonella histaminiformans, Collinsella aerofaciens, and Eubacterium ruminantium* showed insignificant positive overall effect estimates across all 12 studies, while the remaining showed negative effects[Figure 3.18]. In the case of the Jaccard measure, we could see most of the species are negatively associated while species like *Collinsella aerofaciens, Eubacterium hallii, and Ruminococcus torques* were found to have insignificantly overall

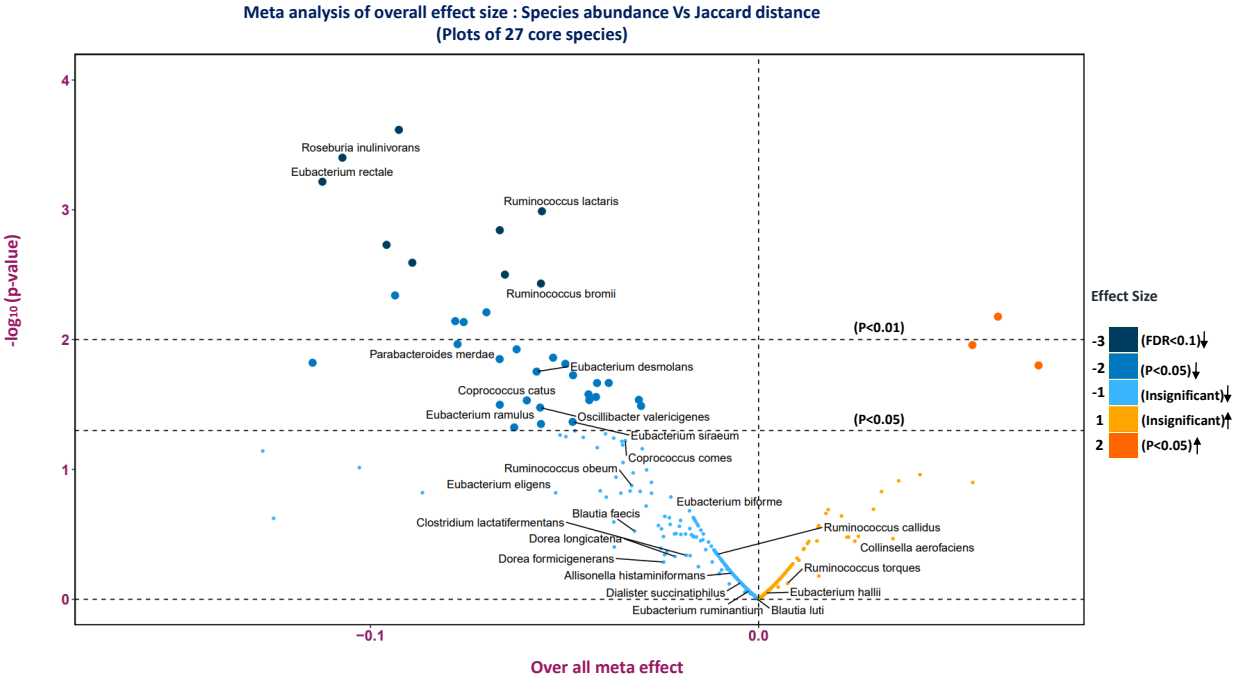

Figure 3.16: Meta effect analysis all species with respect to Jaccard distance (27 species highlighted

positive effect estimates[Figure 3.19]. Finally, overall estimates of the 27 species showed a completely negative effect size while analyzing these cores species' abundance with the Aitchison distance measures[Figure 3.20].

3.2.2 The extent of the estimated negative association between the core gut microbiome and across time-point gut microbiome variability was comparatively high compared to the rest of the species

With the above effect analysis of core gut microbes, we found that most species were negatively correlated with three distance measures. So, we next compared the extent of the negative summarized effects of these core taxa on the longitudinal beta diversity compared to the remaining 313 species[Figure 3.21]. In [Figure 3.21A], a clear significant distinction between the distribution of overall estimates of the core groups and the remaining groups of taxa. The trend of estimates in core groups is towards negative

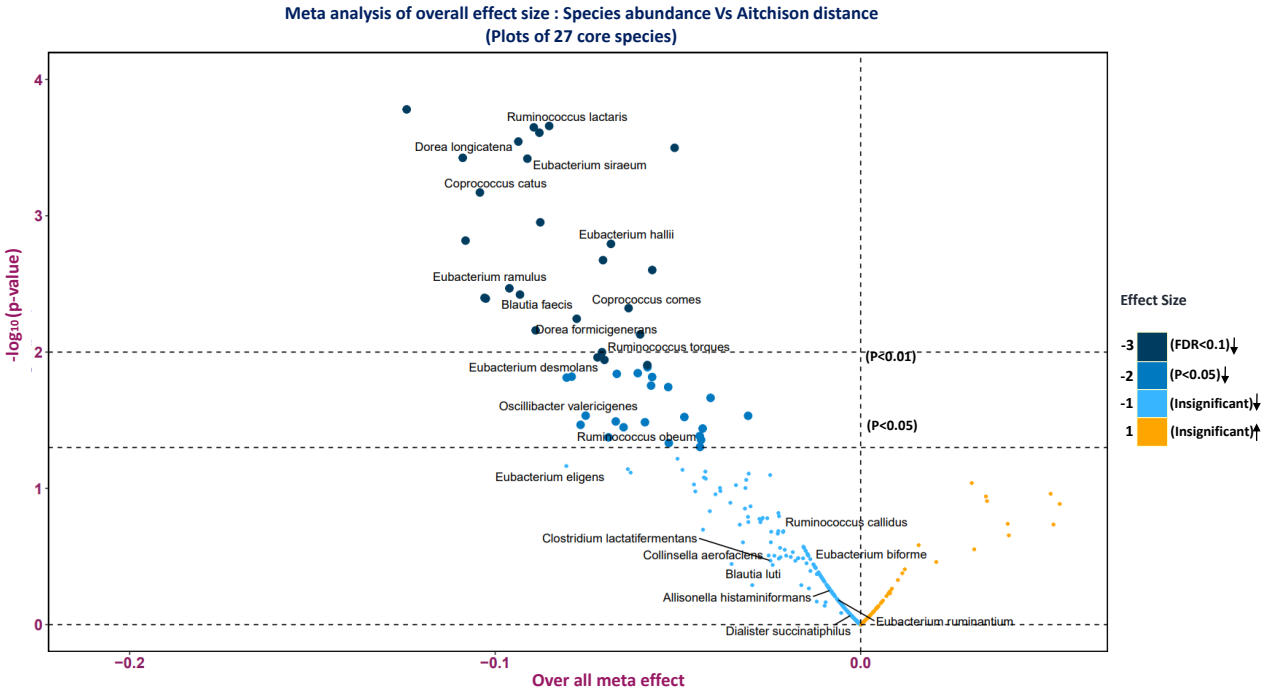

Figure 3.17: Meta effect analysis all species with respect to Aitchison distance (27 species highlighted

values across all three distance measures, while the median estimates of species other than the core were found to be close to zero. This variation in effect across these two groups(core v/s others) was statistically significant, with $p<0.01$. Next, we also compared the proportion of negative effect estimates v/s combined positive and insignificant negative effects in both groups. We could find a significant difference in the proportion of negative effect estimation between these two groups with $p<0.01$. The proportion variation was much more pronounced in the Aitchison distance measure compared to the other two[Figure 3.21B]. The trend of most of these core gut microbiomes showing negative effects on beta diversity across multiple time points shows their significance in maintaining stability in gut microbial communities. The depleted abundance of these core species has a very high negative impact on the beta diversity resulting in very few species with low gut microbiome variability (or high stability or resilience).

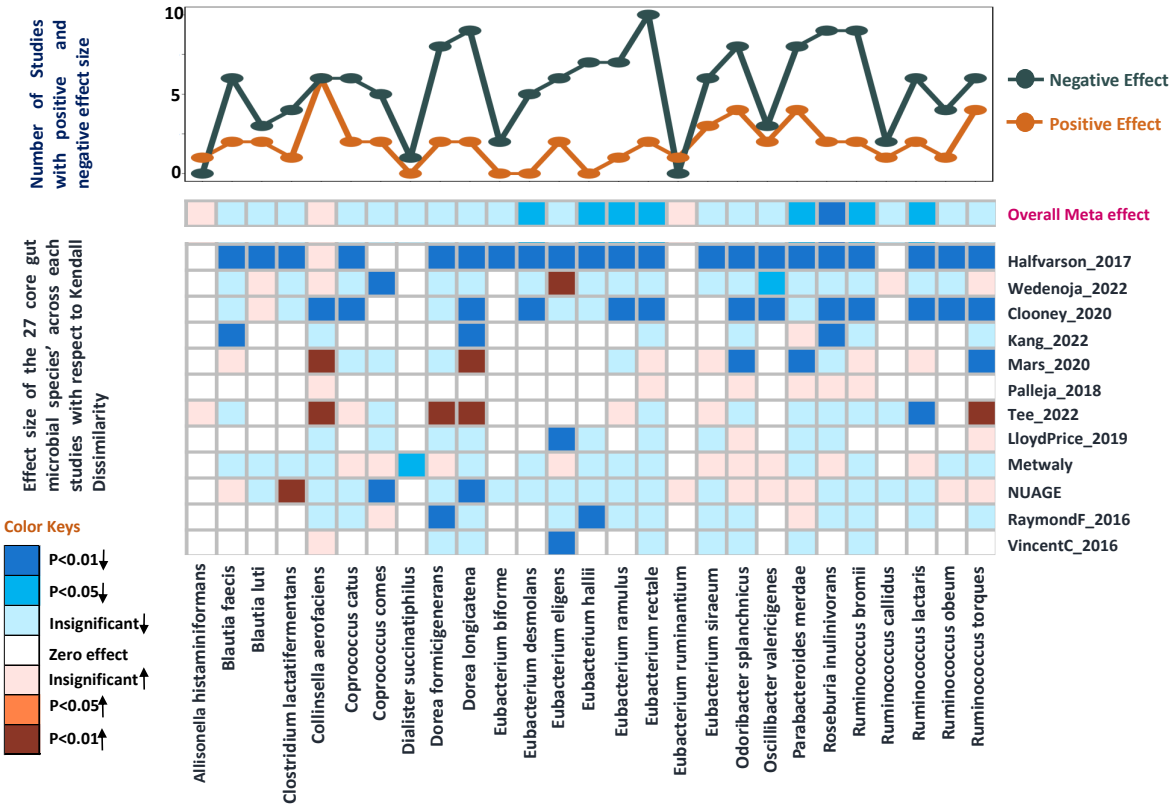

Figure 3.18: Meta effect analysis of core gut species(27) with respect to Kendall distance

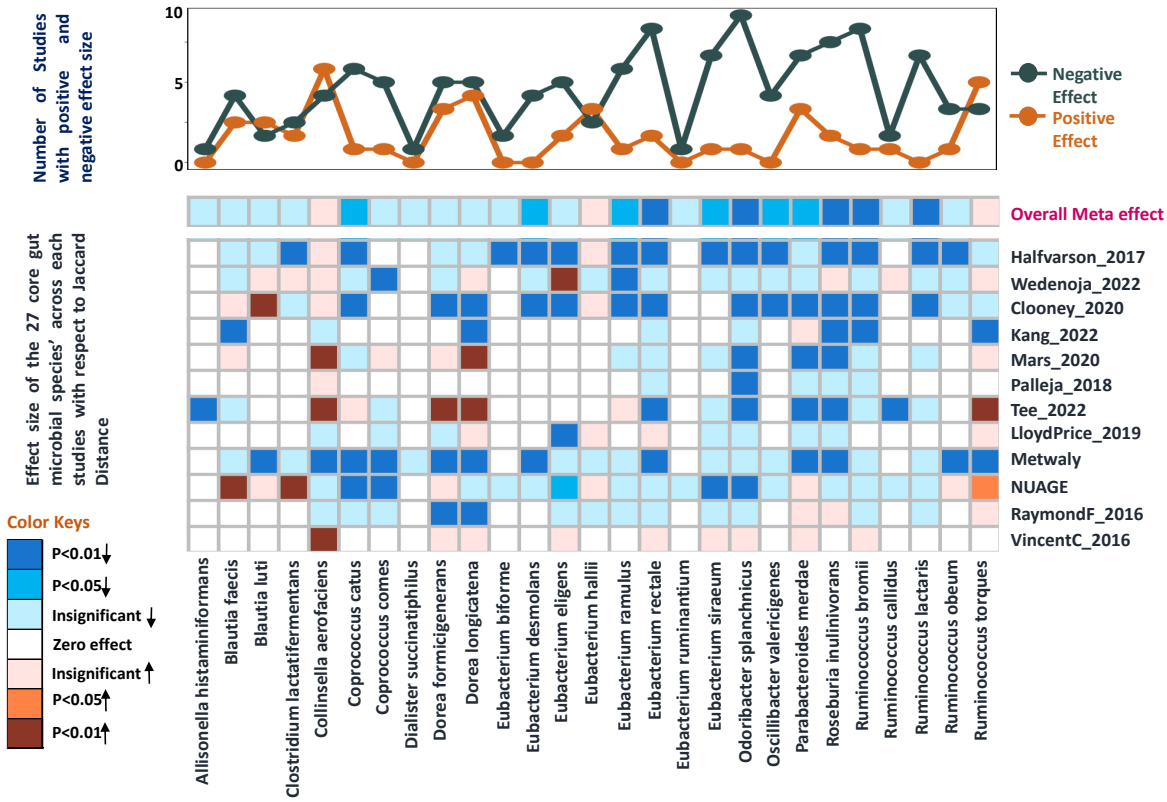

Figure 3.19: Meta effect analysis of core gut species(27) with respect to Jaccard distance

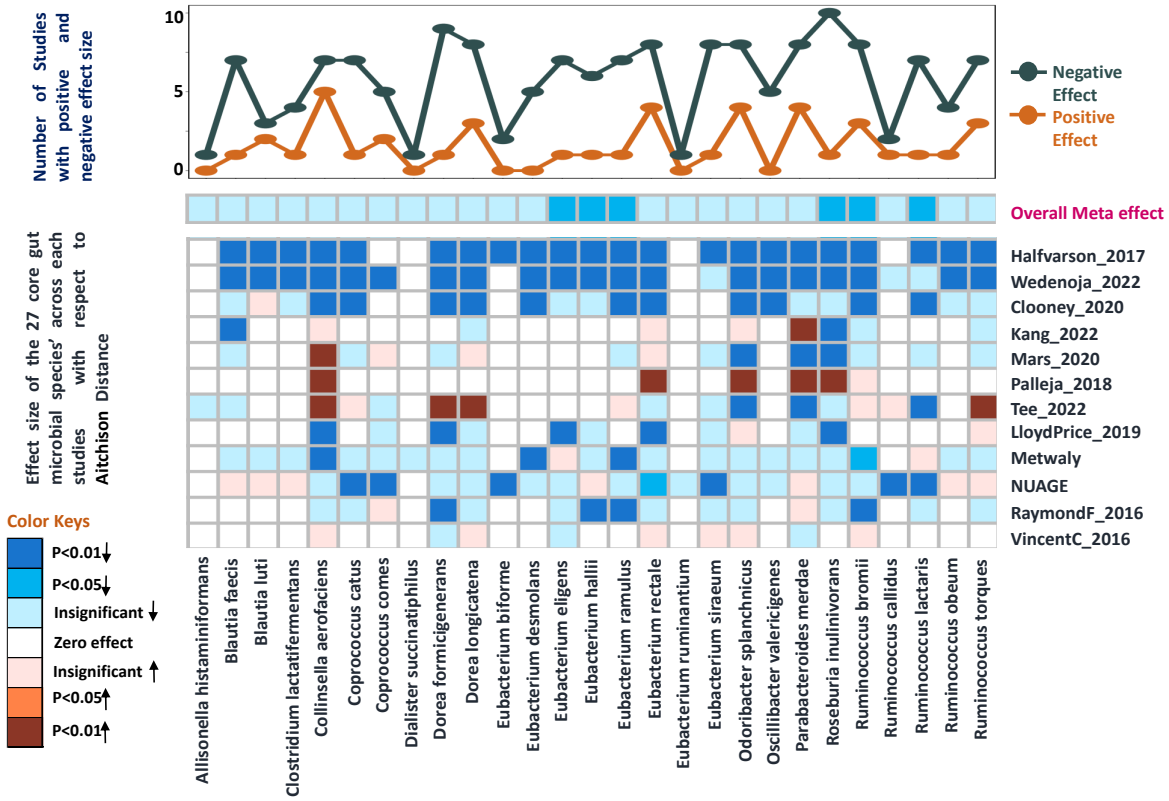

Figure 3.20: Meta effect analysis of core gut species(27) with respect to Aitchison distance

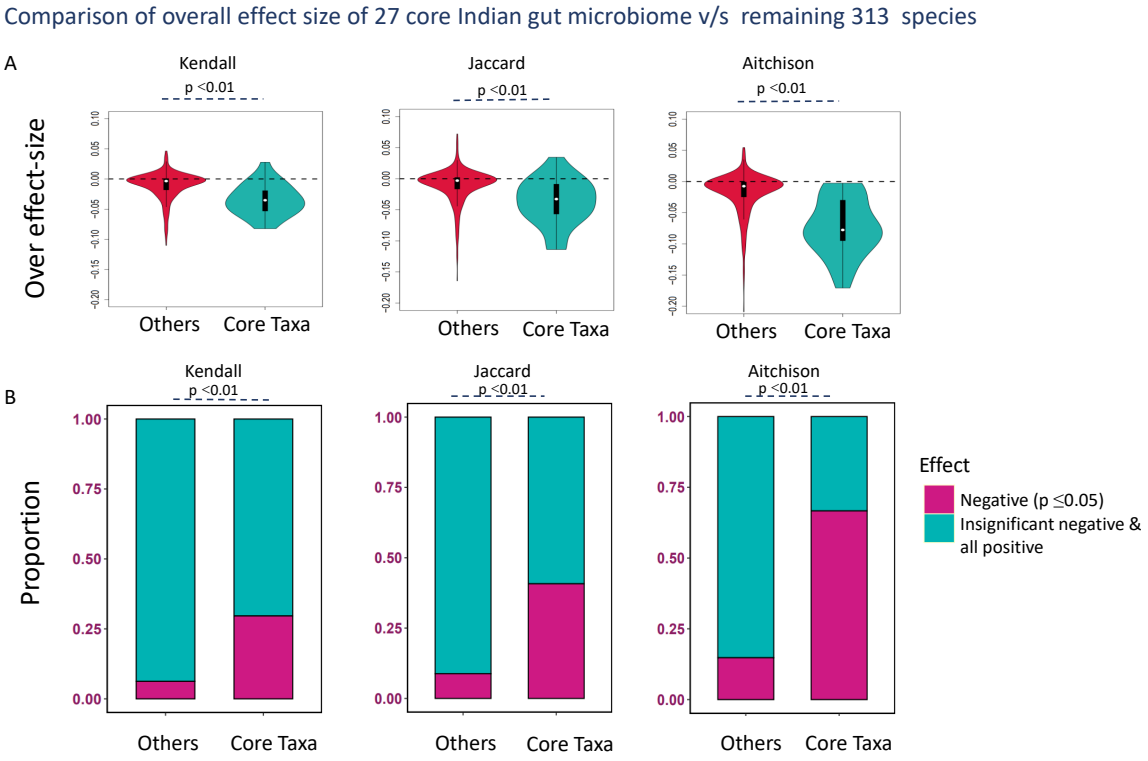

Figure 3.21: Comparison of estimated effects of core taxa v/s others

Chapter 4

Conclusion & Future Scope

4.1 Conclusion

In previous studies, researchers have extensively investigated the variation patterns in gut microbiome communities in cases of Inflammatory Bowel Disease (IBD)[2, 5, 6]. Building upon this knowledge, our study aimed to uncover variations specific to two distinct types of gut inflammatory diseases, Crohn's Disease (CD) and Intestinal Tuberculosis (ITB). Our investigation involved 57 subjects categorized into Controls, CD, and ITB groups. Through alpha diversity analysis using the Shannon Index and Pielou's index, we observed reduced gut microbiome evenness in both CD and ITB groups compared to the controls. Furthermore, ITB exhibited a distinctive reduction in fungal community abundance relative to controls at the mycobiome level, while CD displayed a mycobiome pattern similar to controls. Principal Coordinate Analyses confirmed CD and ITB's distinct gut microbiome and mycobiome compositions.

We also identified taxonomic markers capable of discriminating between the three groups. Moreover, we discovered additional features contributing to disease discrimination using cross-group significantly varying Principal Coordinate Analysis (PCoA) axes. The co-abundance network analysis further divided these diagnostic markers into four modules, each showing unique patterns of alterations. The modules' reproducibility across global studies demonstrated the robustness of disease-related depleted taxa. With the aim of linking these disease-depleted modules to the core gut microbiome, which may be associated with the inflammatory phenotype, we sought to identify the core gut microbiome specific to the Indian population. The core gut microbiome score (CGMS) was calculated based on 46 prevalent and centrally located taxa, consisting of 28 species and 18 genera in the Indian population. The CGMS was a valuable marker for gauging inflammation status in diverse populations.

The core group identified in our study provided valuable insights into deducing a patient's inflammatory status and level of dysbiosis. However, this analysis was limited to a single time point, which may not capture the dynamic changes occurring over time. We extended our analysis to a multi-time point setup to address this limitation and better understand the longitudinal diversity. By observing trends across multiple time points, we could better assess how the core gut taxa varied over time. Interestingly, we found that these core taxa consistently negatively affected distance measures. This discovery further emphasized their crucial role in maintaining stability within the gut microbiome's intra and inter-community networks. The ability of the core gut taxa to consistently influence the distance measures over time underscores their significance in preserving the resilience and balance of gut microbiome communities. Their continued presence and stability offer potential as robust markers for monitoring and assessing the health status of patients longitudinally.

4.2 Future Scope

In the present longitudinal study, we focused on the core group identified specifically from the Indian population. However, in future research, this investigation can be extended by incorporating data from a comprehensive meta-analysis study that includes datasets from diverse populations worldwide. By incorporating a global core microbiome dataset, we can obtain a more generalized and representative understanding of the essential microbial components involved in maintaining the stability and balance of gut microbiome network modules. This expanded approach will enable us to capture the full spectrum of microbial interactions and variations that exist across different geographical regions and ethnicities. Furthermore, including global core microbiome data can enhance the diagnostic potential of these markers in gut inflammation. As we study the core taxa from a broader population perspective, we can identify common patterns and unique variations associated with inflammatory conditions that may overcome regional differences. This can lead to the developing of more robust and universally applicable diagnostic markers for gut inflammation applicable to diverse populations worldwide. Additionally, a global perspective will provide valuable insights into the impact of lifestyle, diet, and environmental factors on the gut microbiome. It will help us understand how these factors interact with the core microbial components to influence gut health and disease outcomes across various populations.

Bibliography

[1] J. M. Pickard, M. Y. Zeng, R. Caruso, and G. Núñez, "Gut Microbiota: Role in Pathogen Colonization, Immune Responses and Inflammatory Disease," *Immunological reviews*, vol. 279, no. 1, pp. 70–89, Sep. 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5657496/

[2] E. A. Franzosa, A. Sirota-Madi, J. Avila-Pacheco, N. Fornelos, H. J. Haiser, S. Reinker, T. Vatanen, A. B. Hall, H. Mallick, L. J. McIver, J. S. Sauk, R. G. Wilson, B. W. Stevens, J. M. Scott, K. Pierce, A. A. Deik, K. Bullock, F. Imhann, J. A. Porter, A. Zhernakova, J. Fu, R. K. Weersma, C. Wijmenga, C. B. Clish, H. Vlamakis, C. Huttenhower, and R. J. Xavier, "Gut microbiome structure and metabolic activity in inflammatory bowel disease," *Nature Microbiology*, vol. 4, no. 2, pp. 293–305, Feb. 2019, number: 2 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41564-018-0306-4

[3] E. H. Crost, E. Coletto, A. Bell, and N. Juge, "Ruminococcus gnavus: friend or foe for human health," *FEMS Microbiology Reviews*, vol. 47, no. 2, p. fuad014, Mar. 2023. [Online]. Available: https://doi.org/10.1093/femsre/fuad014

[4] P. B. Eckburg, E. M. Bik, C. N. Bernstein, E. Purdom, L. Dethlefsen, M. Sargent, S. R. Gill, K. E. Nelson, and D. A. Relman, "Diversity of the human intestinal microbial flora," *Science (New York, N.Y.)*, vol. 308, no. 5728, pp. 1635–1638, Jun. 2005.

[5] J. Lloyd-Price, C. Arze, A. N. Ananthakrishnan, M. Schirmer, J. Avila-Pacheco, T. W. Poon, E. Andrews, N. J. Ajami, K. S. Bonham, C. J. Brislawn, D. Casero, H. Courtney, A. Gonzalez, T. G. Graeber, A. B. Hall, K. Lake, C. J. Landers, H. Mallick, D. R. Plichta, M. Prasad, G. Rahnavard, J. Sauk, D. Shungin, Y. Vázquez-Baeza, R. A. White, J. Braun, L. A. Denson, J. K. Jansson, R. Knight, S. Kugathasan, D. P. B. McGovern, J. F. Petrosino, T. S. Stappenbeck, H. S. Winter, C. B. Clish, E. A. Franzosa, H. Vlamakis, R. J.

Xavier, and C. Huttenhower, "Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases," *Nature*, vol. 569, no. 7758, pp. 655–662, May 2019, number: 7758 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41586-019-1237-9

[6] J. Halfvarson, C. J. Brislawn, R. Lamendella, Y. Vázquez-Baeza, W. A. Walters, L. M. Bramer, M. D'Amato, F. Bonfiglio, D. McDonald, A. Gonzalez, E. E. McClure, M. F. Dunklebarger, R. Knight, and J. K. Jansson, "Dynamics of the human gut microbiome in inflammatory bowel disease," *Nature Microbiology*, vol. 2, no. 5, pp. 1–7, Feb. 2017, number: 5 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/nmicrobiol20174

[7] World Health Organization, *Global tuberculosis report 2015*. World Health Organization, 2015. [Online]. Available: https://apps.who.int/iris/handle/10665/191102

[8] C. E. Barry, H. I. Boshoff, V. Dartois, T. Dick, S. Ehrt, J. Flynn, D. Schnappinger, R. J. Wilkinson, and D. Young, "The spectrum of latent tuberculosis: rethinking the biology and intervention strategies," *Nature Reviews. Microbiology*, vol. 7, no. 12, pp. 845–855, Dec. 2009.

[9] J. B. Marshall, "Tuberculosis of the gastrointestinal tract and peritoneum," *The American Journal of Gastroenterology*, vol. 88, no. 7, pp. 989–999, Jul. 1993.

[10] T. Yamane, A. Umeda, and H. Shimao, "Analysis of recent cases of intestinal tuberculosis in Japan," *Internal Medicine (Tokyo, Japan)*, vol. 53, no. 9, pp. 957–962, 2014.

[11] S. Kedia, P. Das, K. S. Madhusudhan, S. Dattagupta, R. Sharma, P. Sahni, G. Makharia, and V. Ahuja, "Differentiating Crohn's disease from intestinal tuberculosis," *World Journal of Gastroenterology*, vol. 25, no. 4, pp. 418–432, Jan. 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6350172/

[12] A. E. M'Koma, "Inflammatory Bowel Disease: An Expanding Global Health Problem," *Clinical Medicine Insights. Gastroenterology*, vol. 6, pp. 33–47, Aug. 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4020403/

[13] C. He, H. Wang, C. Yu, C. Peng, X. Shu, W. Liao, and Z. Zhu, "Alterations of Gut Microbiota in Patients With Intestinal Tuberculosis That Different

From Crohn's Disease," *Frontiers in Bioengineering and Biotechnology*, vol. 9, p. 673691, Jul. 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8290844/

[14] P. T. Santana, S. L. B. Rosas, B. E. Ribeiro, Y. Marinho, and H. S. P. de Souza, "Dysbiosis in Inflammatory Bowel Disease: Pathogenic Role and Potential Therapeutic Targets," *International Journal of Molecular Sciences*, vol. 23, no. 7, p. 3464, Mar. 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8998182/

[15] S. R. Carding, N. Davis, and L. Hoyles, "Review article: the human intestinal virome in health and disease," *Alimentary Pharmacology & Therapeutics*, vol. 46, no. 9, pp. 800–815, Nov. 2017.

[16] S. Kedia, T. S. Ghosh, S. Jain, A. Desigamani, A. Kumar, V. Gupta, S. Bopanna, D. P. Yadav, S. Goyal, G. Makharia, S. P. L. Travis, B. Das, and V. Ahuja, "Gut microbiome diversity in acute severe colitis is distinct from mild to moderate ulcerative colitis," *Journal of Gastroenterology and Hepatology*, vol. 36, no. 3, pp. 731–739, Mar. 2021.

[17] B. Das, T. S. Ghosh, S. Kedia, R. Rampal, S. Saxena, S. Bag, R. Mitra, M. Dayal, O. Mehta, A. Surendranath, S. P. L. Travis, P. Tripathi, G. B. Nair, and V. Ahuja, "Analysis of the Gut Microbiome of Rural and Urban Healthy Indians Living in Sea Level and High Altitude Areas," *Scientific Reports*, vol. 8, no. 1, p. 10104, Jul. 2018, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41598-018-28550-3

[18] N. K. Pinna, R. M. Anjana, S. Saxena, A. Dutta, V. Gnanaprakash, G. Rameshkumar, S. Aswath, S. Raghavan, C. S. S. Rani, V. Radha, M. Balasubramanyam, A. Pant, T. Nielsen, T. Jørgensen, K. Færch, A. Kashani, M. C. A. Silva, H. Vestergaard, T. H. Hansen, T. Hansen, M. Arumugam, G. B. Nair, B. Das, O. Pedersen, V. Mohan, and S. S. Mande, "Trans-ethnic gut microbial signatures of prediabetic subjects from India and Denmark," *Genome Medicine*, vol. 13, no. 1, p. 36, Mar. 2021. [Online]. Available: https://doi.org/10.1186/s13073-021-00851-9

[19] C. Alvarez-Silva, A. Kashani, T. H. Hansen, N. K. Pinna, R. M. Anjana, A. Dutta, S. Saxena, J. Støy, U. Kampmann, T. Nielsen, T. Jørgensen, V. Gnanaprakash, R. Gnanavadivel, A. Sukumaran, C. S. S. Rani, K. Færch, V. Radha, M. Balasubramanyam, G. B. Nair, B. Das, H. Vestergaard,

T. Hansen, S. S. Mande, V. Mohan, M. Arumugam, and O. Pedersen, "Trans-ethnic gut microbiota signatures of type 2 diabetes in Denmark and India," *Genome Medicine*, vol. 13, no. 1, p. 37, Mar. 2021. [Online]. Available: https://doi.org/10.1186/s13073-021-00856-4

[20] A. K. Dubey, N. Uppadhyaya, P. Nilawe, N. Chauhan, S. Kumar, U. A. Gupta, and A. Bhaduri, "LogMPIE, pan-India profiling of the human gut microbiome using 16S rRNA sequencing," *Scientific Data*, vol. 5, no. 1, p. 180232, Oct. 2018, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/sdata2018232

[21] D. B. Dhakan, A. Maji, A. K. Sharma, R. Saxena, J. Pulikkan, T. Grace, A. Gomez, J. Scaria, K. R. Amato, and V. K. Sharma, "The unique composition of Indian gut microbiome, gene catalogue, and associated fecal metabolome deciphered using multi-omics approaches," *GigaScience*, vol. 8, no. 3, p. giz004, Mar. 2019.

[22] A. Gupta, D. B. Dhakan, A. Maji, R. Saxena, V. P. P K, S. Mahajan, J. Pulikkan, J. Kurian, A. M. Gomez, J. Scaria, K. R. Amato, A. K. Sharma, and V. K. Sharma, "Association of Flavonifractor plautii, a Flavonoid-Degrading Bacterium, with the Gut Microbiome of Colorectal Cancer Patients in India," *mSystems*, vol. 4, no. 6, pp. e00 438–19, Nov. 2019.

[23] S. Pareek, T. Kurakawa, B. Das, D. Motooka, S. Nakaya, T. Rongsen-Chandola, N. Goyal, H. Kayama, D. Dodd, R. Okumura, Y. Maeda, K. Fujimoto, T. Nii, T. Ogawa, T. Iida, N. Bhandari, T. Kida, S. Nakamura, G. B. Nair, and K. Takeda, "Comparison of Japanese and Indian intestinal microbiota shows diet-dependent interaction between bacteria and fungi," *npj Biofilms and Microbiomes*, vol. 5, no. 1, pp. 1–13, Dec. 2019, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41522-019-0110-9

[24] G. Liguori, B. Lamas, M. L. Richard, G. Brandi, G. da Costa, T. W. Hoffmann, M. P. Di Simone, C. Calabrese, G. Poggioli, P. Langella, M. Campieri, and H. Sokol, "Fungal Dysbiosis in Mucosa-associated Microbiota of Crohn's Disease Patients," *Journal of Crohn's & Colitis*, vol. 10, no. 3, pp. 296–305, Mar. 2016.

[25] I. B. Jeffery, A. Das, E. O'Herlihy, S. Coughlan, K. Cisek, M. Moore, F. Bradley, T. Carty, M. Pradhan, C. Dwibedi, F. Shanahan, and P. W. O'Toole, "Differences in Fecal Microbiomes and Metabolomes of People With vs Without Irritable Bowel Syndrome and Bile Acid Malabsorption," *Gastroenterology*, vol. 158, no. 4, pp. 1016–1028.e8, Mar. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0016508519419203

[26] A. Das, E. O'Herlihy, F. Shanahan, P. W. O'Toole, and I. B. Jeffery, "The fecal mycobiome in patients with Irritable Bowel Syndrome," *Scientific Reports*, vol. 11, no. 1, p. 124, Jan. 2021.

[27] A. K. Nash, T. A. Auchtung, M. C. Wong, D. P. Smith, J. R. Gesell, M. C. Ross, C. J. Stewart, G. A. Metcalf, D. M. Muzny, R. A. Gibbs, N. J. Ajami, and J. F. Petrosino, "The gut mycobiome of the Human Microbiome Project healthy cohort," *Microbiome*, vol. 5, no. 1, p. 153, Nov. 2017. [Online]. Available: https://doi.org/10.1186/s40168-017-0373-4

[28] Yong Yang, Z. Han, Z. Gao, J. Chen, C. Song, J. Xu, H. Wang, A. Huang, J. Shi, and J. Gu, "Metagenomic and targeted metabolomic analyses reveal distinct phenotypes of the gut microbiota in patients with colorectal cancer and type 2 diabetes mellitus," *Chinese Medical Journal*, p. 10.1097/CM9.0000000000002421. [Online]. Available: https://journals.lww.com/cmj/Fulltext/9900/Metagenomic_and_targeted_metabolomic_analyses.472.aspx

[29] J. Hu, S. Wei, Y. Gu, Y. Wang, Y. Feng, J. Sheng, L. Hu, C. Gu, P. Jiang, Y. Tian, W. Guo, L. Lv, F. Liu, Y. Zou, F. Yan, and N. Feng, "Gut Mycobiome in Patients With Chronic Kidney Disease Was Altered and Associated With Immunological Profiles," *Frontiers in Immunology*, vol. 13, p. 843695, Jun. 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9245424/

[30] Y. He, W. Wu, H.-M. Zheng, P. Li, D. McDonald, H.-F. Sheng, M.-X. Chen, Z.-H. Chen, G.-Y. Ji, Z.-D.-X. Zheng, P. Mujagond, X.-J. Chen, Z.-H. Rong, P. Chen, L.-Y. Lyu, X. Wang, C.-B. Wu, N. Yu, Y.-J. Xu, J. Yin, J. Raes, R. Knight, W.-J. Ma, and H.-W. Zhou, "Regional variation limits applications of healthy gut microbiome reference ranges and disease models," *Nature Medicine*, vol. 24, no. 10, pp. 1532–1535, Oct. 2018.

[31] R. Jayasudha, T. Das, S. Kalyana Chakravarthy, G. Sai Prashanthi, A. Bhargava, M. Tyagi, P. K. Rani, R. R. Pappuru, and S. Shivaji, "Gut mycobiomes are altered in people with type 2 Diabetes Mellitus and Diabetic Retinopathy," *PloS One*, vol. 15, no. 12, p. e0243077, 2020.

[32] A. B. Hall, M. Yassour, J. Sauk, A. Garner, X. Jiang, T. Arthur, G. K. Lagoudas, T. Vatanen, N. Fornelos, R. Wilson, M. Bertha, M. Cohen, J. Garber, H. Khalili, D. Gevers, A. N. Ananthakrishnan, S. Kugathasan, E. S. Lander, P. Blainey, H. Vlamakis, R. J. Xavier, and C. Huttenhower, "A novel Ruminococcus gnavus clade enriched in inflammatory bowel disease

patients," *Genome Medicine*, vol. 9, no. 1, p. 103, Nov. 2017. [Online]. Available: https://doi.org/10.1186/s13073-017-0490-5

[33] S. Wedenoja, A. Saarikivi, J. Mälkönen, S. Leskinen, M. Lehto, K. Adeshara, J. Tuokkola, A. Nikkonen, L. Merras-Salmio, M. Höyhtyä, S. Hörkkö, A. Haaramo, A. Salonen, W. M. d. Vos, K. Korpela, and K.-L. Kolho, "Fecal microbiota in congenital chloride diarrhea and inflammatory bowel disease," *PLOS ONE*, vol. 17, no. 6, p. e0269561, Jun. 2022, publisher: Public Library of Science. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0269561

[34] M. Rat, Y. Y, W. T, H. M, P. S, L. Hr, T. X, S. Z, K. Kr, K. T, B. Y, Z. T, B. N, F. G, J. Aj, v. T. W, H. S, O. T, G. M, S. J, D. M, C. M, E. E, S. E, B. R, F. G, S. Jr, K. D, and K. Pc, "Longitudinal Multi-omics Reveals Subset-Specific Mechanisms Underlying Irritable Bowel Syndrome," *Cell*, vol. 182, no. 6, Sep. 2020, publisher: Cell. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/32916129/

[35] A. Palleja, K. H. Mikkelsen, S. K. Forslund, A. Kashani, K. H. Allin, T. Nielsen, T. H. Hansen, S. Liang, Q. Feng, C. Zhang, P. T. Pyl, L. P. Coelho, H. Yang, J. Wang, A. Typas, M. F. Nielsen, H. B. Nielsen, P. Bork, J. Wang, T. Vilsbøll, T. Hansen, F. K. Knop, M. Arumugam, and O. Pedersen, "Recovery of gut microbiota of healthy adults following antibiotic exposure," *Nature microbiology*, vol. 3, no. 11, pp. 1255–1265, Nov. 2018. [Online]. Available: https://doi.org/10.1038/s41564-018-0257-9

[36] J. T. L. Kang, J. J. Y. Teo, D. Bertrand, A. Ng, A. Ravikrishnan, M. Yong, O. T. Ng, K. Marimuthu, S. L. Chen, K. R. Chng, Y.-H. Gan, and N. Nagarajan, "Long-term ecological and evolutionary dynamics in the gut microbiomes of carbapenemase-producing Enterobacteriaceae colonized subjects," *Nature Microbiology*, vol. 7, no. 10, pp. 1516–1524, Oct. 2022, number: 10 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41564-022-01221-w

[37] A. G. Clooney, J. Eckenberger, E. Laserna-Mendieta, K. A. Sexton, M. T. Bernstein, K. Vagianos, M. Sargent, F. J. Ryan, C. Moran, D. Sheehan, R. D. Sleator, L. E. Targownik, C. N. Bernstein, F. Shanahan, and M. J. Claesson, "Ranking microbiome variance in inflammatory bowel disease: a large longitudinal intercontinental study," *Gut*, vol. 70, no. 3, pp. 499–510, Mar. 2021.

[38] M. Z. Tee, Y. X. Er, A. V. Easton, N. J. Yap, I. L. Lee, J. Devlin, Z. Chen, K. S. Ng, P. Subramanian, A. Angelova, O. Oyesola, S. Sargsian, R. Ngui, D. P. Beiting, C. C. M. Boey, K. H. Chua, K. Cadwell, Y. A. L. Lim, P. Loke, and S. C. Lee, "Gut microbiome of helminth-infected indigenous Malaysians is context dependent," *Microbiome*, vol. 10, no. 1, p. 214, Dec. 2022.

[39] A. Metwaly, A. Dunkel, N. Waldschmitt, A. C. D. Raj, I. Lagkouvardos, A. M. Corraliza, A. Mayorgas, M. Martinez-Medina, S. Reiter, M. Schloter, T. Hofmann, M. Allez, J. Panes, A. Salas, and D. Haller, "Integrated microbiota and metabolite profiles link Crohn's disease to sulfur metabolism," *Nature Communications*, vol. 11, no. 1, p. 4322, Aug. 2020, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41467-020-17956-1

[40] T. S. Ghosh, S. Rampelli, I. B. Jeffery, A. Santoro, M. Neto, M. Capri, E. Giampieri, A. Jennings, M. Candela, S. Turroni, E. G. Zoetendal, G. D. A. Hermes, C. Elodie, N. Meunier, C. M. Brugere, E. Pujos-Guillot, A. M. Berendsen, L. C. P. G. M. D. Groot, E. J. M. Feskins, J. Kaluza, B. Pietruszka, M. J. Bielak, B. Comte, M. Maijo-Ferre, C. Nicoletti, W. M. D. Vos, S. Fairweather-Tait, A. Cassidy, P. Brigidi, C. Franceschi, and P. W. O'Toole, "Mediterranean diet intervention alters the gut microbiome in older people reducing frailty and improving health status: the NU-AGE 1-year dietary intervention across five European countries," *Gut*, vol. 69, no. 7, pp. 1218–1228, Jul. 2020, publisher: BMJ Publishing Group Section: Nutrition. [Online]. Available: https://gut.bmj.com/content/69/7/1218

[41] F. Raymond, A. A. Ouameur, M. Déraspe, N. Iqbal, H. Gingras, B. Dridi, P. Leprohon, P.-L. Plante, R. Giroux, Bérubé, J. Frenette, D. K. Boudreau, J.-L. Simard, I. Chabot, M.-C. Domingo, S. Trottier, M. Boissinot, A. Huletsky, P. H. Roy, M. Ouellette, M. G. Bergeron, and J. Corbeil, "The initial state of the human gut microbiome determines its reshaping by antibiotics," *The ISME Journal*, vol. 10, no. 3, pp. 707–720, Mar. 2016, number: 3 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/ismej2015148

[42] C. Vincent, M. A. Miller, T. J. Edens, S. Mehrotra, K. Dewar, and A. R. Manges, "Bloom and bust: intestinal microbiota dynamics in response to hospital exposures and Clostridium difficile colonization or infection," *Microbiome*, vol. 4, no. 1, pp. 1–11, Dec. 2016, number: 1 Publisher: BioMed Central. [Online]. Available: https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-016-0156-3

[43] E. Bolyen, J. R. Rideout, M. R. Dillon, N. A. Bokulich, C. C. Abnet, G. A.

Al-Ghalith, H. Alexander, E. J. Alm, M. Arumugam, F. Asnicar, Y. Bai, J. E. Bisanz, K. Bittinger, A. Brejnrod, C. J. Brislawn, C. T. Brown, B. J. Callahan, A. M. Caraballo-Rodríguez, J. Chase, E. K. Cope, R. Da Silva, C. Diener, P. C. Dorrestein, G. M. Douglas, D. M. Durall, C. Duvallet, C. F. Edwardson, M. Ernst, M. Estaki, J. Fouquier, J. M. Gauglitz, S. M. Gibbons, D. L. Gibson, A. Gonzalez, K. Gorlick, J. Guo, B. Hillmann, S. Holmes, H. Holste, C. Huttenhower, G. A. Huttley, S. Janssen, A. K. Jarmusch, L. Jiang, B. D. Kaehler, K. B. Kang, C. R. Keefe, P. Keim, S. T. Kelley, D. Knights, I. Koester, T. Kosciolek, J. Kreps, M. G. I. Langille, J. Lee, R. Ley, Y.-X. Liu, E. Loftfield, C. Lozupone, M. Maher, C. Marotz, B. D. Martin, D. McDonald, L. J. McIver, A. V. Melnik, J. L. Metcalf, S. C. Morgan, J. T. Morton, A. T. Naimey, J. A. Navas-Molina, L. F. Nothias, S. B. Orchanian, T. Pearson, S. L. Peoples, D. Petras, M. L. Preuss, E. Pruesse, L. B. Rasmussen, A. Rivers, M. S. Robeson, P. Rosenthal, N. Segata, M. Shaffer, A. Shiffer, R. Sinha, S. J. Song, J. R. Spear, A. D. Swafford, L. R. Thompson, P. J. Torres, P. Trinh, A. Tripathi, P. J. Turnbaugh, S. Ul-Hasan, J. J. J. van der Hooft, F. Vargas, Y. Vázquez-Baeza, E. Vogtmann, M. von Hippel, W. Walters, Y. Wan, M. Wang, J. Warren, K. C. Weber, C. H. D. Williamson, A. D. Willis, Z. Z. Xu, J. R. Zaneveld, Y. Zhang, Q. Zhu, R. Knight, and J. G. Caporaso, "Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2," *Nature Biotechnology*, vol. 37, no. 8, pp. 852–857, Aug. 2019, number: 8 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41587-019-0209-9

[44] B. J. Callahan, P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson, and S. P. Holmes, "DADA2: High-resolution sample inference from Illumina amplicon data," *Nature Methods*, vol. 13, no. 7, pp. 581–583, Jul. 2016, number: 7 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/nmeth.3869

[45] G. Allard, F. J. Ryan, I. B. Jeffery, and M. J. Claesson, "SPINGO: a rapid species-classifier for microbial amplicon sequences," *BMC Bioinformatics*, vol. 16, no. 1, p. 324, Oct. 2015. [Online]. Available: https://doi.org/10.1186/s12859-015-0747-1

[46] T. Rognes, T. Flouri, B. Nichols, C. Quince, and F. Mahé, "VSEARCH: a versatile open source tool for metagenomics," *PeerJ*, vol. 4, p. e2584, 2016.

[47] Ghsosh et al., "Toward an improved definition of a healthy microbiome for healthy aging | Nature Aging." [Online]. Available: https://www.nature.com/articles/s43587-022-00306-9

[48] K. Faust, J. F. Sathirapongsasuti, J. Izard, N. Segata, D. Gevers, J. Raes, and C. Huttenhower, "Microbial Co-occurrence Relationships in the Human Microbiome," *PLOS Computational Biology*, vol. 8, no. 7, p. e1002606, Jul. 2012, publisher: Public Library of Science. [Online]. Available: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002606

[49] A. Blanco-Míguez, F. Beghini, F. Cumbo, L. J. McIver, K. N. Thompson, M. Zolfo, P. Manghi, L. Dubois, K. D. Huang, A. M. Thomas, W. A. Nickols, G. Piccinno, E. Piperni, M. Punčochář, M. Valles-Colomer, A. Tett, F. Giordano, R. Davies, J. Wolf, S. E. Berry, T. D. Spector, E. A. Franzosa, E. Pasolli, F. Asnicar, C. Huttenhower, and N. Segata, "Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4," *Nature Biotechnology*, pp. 1–12, Feb. 2023, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41587-023-01688-w

[50] G. Hoarau, P. K. Mukherjee, C. Gower-Rousseau, C. Hager, J. Chandra, M. A. Retuerto, C. Neut, S. Vermeire, J. Clemente, J. F. Colombel, H. Fujioka, D. Poulain, B. Sendid, and M. A. Ghannoum, "Bacteriome and Mycobiome Interactions Underscore Microbial Dysbiosis in Familial Crohn's Disease," *mBio*, vol. 7, no. 5, pp. 10.1128/mbio.01 250–16, Sep. 2016, publisher: American Society for Microbiology. [Online]. Available: https://journals.asm.org/doi/full/10.1128/mBio.01250-16

[51] H. Chu, A. Khosravi, I. P. Kusumawardhani, A. H. K. Kwon, A. C. Vasconcelos, L. D. Cunha, A. E. Mayer, Y. Shen, W.-L. Wu, A. Kambal, S. R. Targan, R. J. Xavier, P. B. Ernst, D. R. Green, D. P. B. McGovern, H. W. Virgin, and S. K. Mazmanian, "Gene-microbiota interactions contribute to the pathogenesis of inflammatory bowel disease," *Science (New York, N.Y.)*, vol. 352, no. 6289, pp. 1116–1120, May 2016.

[52] M. Schirmer, L. Denson, H. Vlamakis, E. A. Franzosa, S. Thomas, N. M. Gotman, P. Rufo, S. S. Baker, C. Sauer, J. Markowitz, M. Pfefferkorn, M. Oliva-Hemker, J. Rosh, A. Otley, B. Boyle, D. Mack, R. Baldassano, D. Keljo, N. LeLeiko, M. Heyman, A. Griffiths, A. S. Patel, J. Noe, S. Kugathasan, T. Walters, C. Huttenhower, J. Hyams, and R. J. Xavier, "Compositional and Temporal Changes in the Gut Microbiome of Pediatric Ulcerative Colitis Patients Are Linked to Disease Course," *Cell Host & Microbe*, vol. 24, no. 4, pp. 600–610.e4, Oct. 2018.

[53] D. Gevers, S. Kugathasan, L. A. Denson, Y. Vázquez-Baeza, W. Van Treuren, B. Ren, E. Schwager, D. Knights, S. J. Song, M. Yassour, X. C. Morgan, A. D.

Kostic, C. Luo, A. González, D. McDonald, Y. Haberman, T. Walters, S. Baker, J. Rosh, M. Stephens, M. Heyman, J. Markowitz, R. Baldassano, A. Griffiths, F. Sylvester, D. Mack, S. Kim, W. Crandall, J. Hyams, C. Huttenhower, R. Knight, and R. J. Xavier, "The treatment-naive microbiome in new-onset Crohn's disease," *Cell Host & Microbe*, vol. 15, no. 3, pp. 382–392, Mar. 2014.

[54] V. Pascal, M. Pozuelo, N. Borruel, F. Casellas, D. Campos, A. Santiago, X. Martinez, E. Varela, G. Sarrabayrouse, K. Machiels, S. Vermeire, H. Sokol, F. Guarner, and C. Manichanh, "A microbial signature for Crohn's disease," *Gut*, vol. 66, no. 5, pp. 813–822, May 2017.