CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

Identification of Taxonomic Biomarkers and Multivariable Associations Analysis for Crohn's Disease

by

Abeera

A thesis submitted in partial fulfillment for the degree of Master of Science

in the

Faculty of Health and Life Sciences
Department of Bioinformatics and Biosciences

2025

Dedicated to my parents, husband and daughter.

CERTIFICATE OF APPROVAL

Identification of Taxonomic Biomarkers and Multivariable Associations Analysis for Crohn's Disease

by

Abeera

(MBS233007)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Pakeeza Arzoo Shaiq	PMAS-AAU, Rwp
(b)	Internal Examiner	Dr. Arshia Amin Butt	CUST, Islamabad
(c)	Supervisor	Dr. Syeda Marriam Bakhtiar	CUST, Islamabad

Dr. Syeda Marriam Bakhtiar
Thesis Supervisor
August, 2025

Dr. Syeda Marriam Bakhtiar
Head
Dept. of Bioinfo. & Biosciences
August, 2025

Dr. Sahar Fazal
Dean
Faculty of Health & Life Sciences
August, 2025

Author's Declaration

I, **Abeera** hereby state that my MS thesis titled "**Identification of Taxonomic Biomarkers and Multivariable Associations Analysis for Crohn's Disease**" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.

(Abeera)

Registration No: MBS233007

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled "**Identification of Taxonomic Biomarkers and Multivariable Associations Analysis for Crohn's Disease**" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.

(Abeera)

Registration No: MBS233007

Acknowledgement

Above all, I am thankful to **Almighty Allah** for blessing me with the strength, resilience, and wisdom to complete this milestone.

First and foremost, I express my deepest gratitude to my beloved husband, **Umer Raza**, for his unwavering support, encouragement, and understanding throughout my academic journey. His patience, sacrifices, and belief in my potential have been the cornerstone of my success.

I am equally grateful to my precious daughter, **Zimal Fatima**, whose innocent smiles and unconditional love have been a constant source of motivation and joy during even the most challenging times.

I extend heartfelt thanks to my **parents** for their endless prayers, love, and moral support that have shaped who I am today. To my **siblings**, I owe a debt of gratitude for their encouragement, assistance, and companionship at every step of this journey.

I am also sincerely thankful to my **in-laws** for their understanding and support, which eased the demands of balancing academic and family responsibilities.

A special thanks to my **friends and classmates**, who made this journey less stressful and more memorable through their companionship, advice, and encouragement.

I wish to acknowledge and express my sincere appreciation to all my teachers and mentors, particularly **Dr. Syeda Mariam Bakhtiar**, whose invaluable guidance, constructive feedback, and continuous encouragement made this research possible.

Lastly, I am grateful to the **Department of Bioinformatics and Biosciences, Capital University of Science and Technology**, for providing the academic environment and resources necessary to accomplish my goals. I am grateful for Dean Faculty of Health and Life sciences **Dr Sahar Fazal** and her office for facilitation and guidance throughout my journey. I also appreciate the efforts of Graduate Studies office for making things streamlined for us and enabling timely completion of tasks.

(Abeera)

Abstract

Crohn's disease is a multifactorial complex disorder with various risk factors, where gut microbiome is one of the key contributors in the onset of disease and severity in symptoms. Metagenomic analysis of patients suffering from Crohn's disease reveals that gut microbiota variations contribute to gut dysbiosis. This study was designed to identify the taxonomic biomarkers, i.e., microbial taxa, which could be used as an indicator of disease. After multivariable regression analysis of microbial taxa associations with clinical variables, this study identifies novel microbial signatures strongly linked to CD, including *Alistipes indistinctus* (coef = 1.00, qval = 4.58e-60), *Prevotella copri* (coef = 1.00, qval = 4.62e-60), and *Fusobacterium nucleatum* (coef = 1.00, qval = 2.26e-59), which are significantly enriched in CD patients compared to healthy controls (Diagnosis.pdf). Strikingly, immunosuppressant and steroid therapies drive distinct microbial shifts, suppressing inflammation-associated taxa (e.g., *Blautia pseudococcoides*, coef = -0.42, qval = 8.26e-60) while promoting potential beneficial species (e.g., *Phocaeicola dorei*, coef = 1.65, qval = 0.00016 with mesalamine; mesalamine.pdf). Age-related dynamics further highlight taxa such as *Parabacteroides merdae* (coef = -0.46, qval = 0.063) and *Blautia hansenii* (coef = -0.26, qval = 0.063), which decline with aging (Age.pdf). Statistical rigor was ensured via FDR correction (qval < 0.1), with most associations exhibiting ultra-low q-values (<1e-50), underscoring the robustness of findings. A heatmap visualization (heatmap.pdf) integrates these results, revealing clusters of taxa with shared responses to clinical variables, such as CD-enriched Alistipes and immunosuppressant-depleted Clostridium.

Contents

List of Figures

List of Tables

Abbreviations

ATG16L1	Autophagy Related 16 Like 1
CD	Crohn's Disease
GI	Gastrointestinal
IBD	Inflammatory Bowel Disease
IL23R	Interleukin 23 Receptor
NOD2	Nucleotide-binding Oligomerization Domain-containing protein 2
SCFA	Short-Chain Fatty Acid
SCFAs	Short-Chain Fatty Acids
TNF-α	Tumor Necrosis Factor-alpha

Chapter 1

Introduction

Crohn's disease (CD) is a chronic inflammatory disorder of the gastrointestinal (GI) tract that affects millions worldwide, with a rising prevalence in both developed and developing nations. It is classified as Inflammatory Bowel Disease (IBD) and is characterized by transmural inflammation, leading to complications such as strictures, fistulas, and intestinal damage.

In the past two decades, the global incidence of CD has increased significantly, with North America and Europe having the highest prevalence rates, where approximately 400 per 100,000 individuals are affected. Meanwhile, recent epidemiological studies indicate that the burden of CD is rising in Asian and Middle Eastern countries, including Pakistan, where the prevalence is estimated to be increasing due to rapid urbanization and dietary transitions toward Westernized food patterns, or more precisely, the processed food with preservatives [1, 2]. Despite advancements in diagnosis and treatment, the exact cause of CD remains unknown, and managing its symptoms remains a significant challenge in gastroenterology.

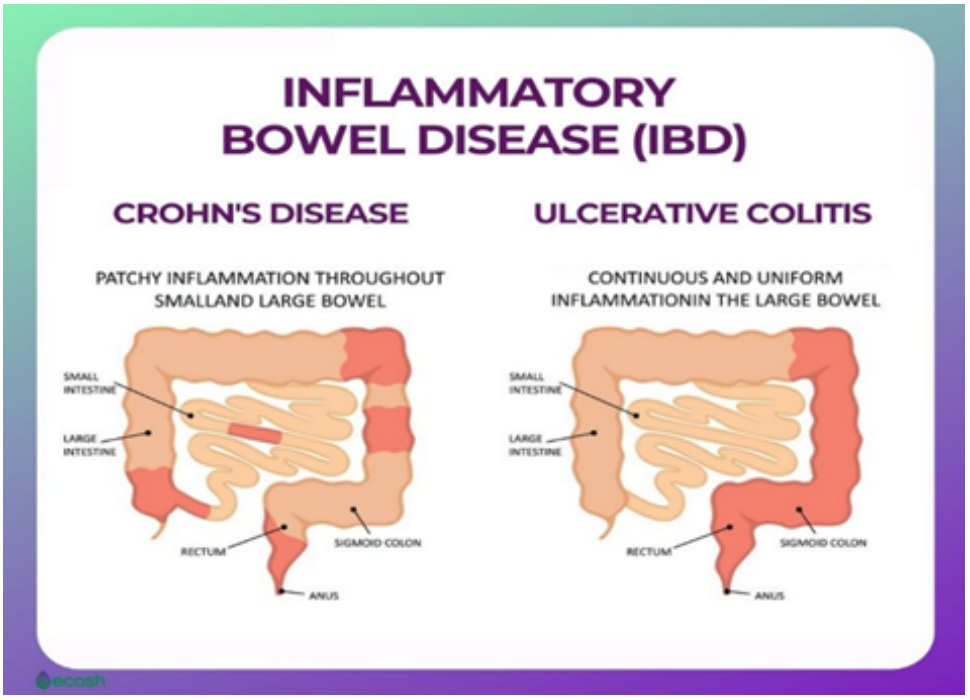

FIGURE 1.1: The classification of inflammatory Bowel Disease into Ulcerative Colitis and Crohn's Disease

In Pakistan, CD has been historically underdiagnosed due to the overlap of symptoms with tuberculosis and other gastrointestinal infections. However, hospital-based reports indicate an increasing trend in the number of diagnosed cases, particularly among young adults between the ages of 15 and 35. A retrospective analysis of IBD cases in Pakistan has revealed that nearly 20% of inflammatory bowel disease patients present with Crohn's disease, suggesting a considerable disease burden. The rise in CD incidence in developing countries has been attributed to environmental shifts, including changes in diet, antibiotic overuse, and industrialization [3, 4]. The lack of population-based registries makes it difficult to estimate the exact prevalence of CD in Pakistan, but clinical studies suggest an alarming rise in cases, requiring further research and healthcare planning to improve early diagnosis and management strategies [5].

The treatment options for Crohn's disease aim to induce and maintain remission, reduce inflammation, and improve the quality of life for patients. Current therapies include anti-inflammatory drugs, corticosteroids, immunosuppressants, and biologics such as tumor necrosis factor-alpha (TNF-α) inhibitors. Among them,

biologic therapies such as infliximab and adalimumab have revolutionized CD management by targeting specific immune pathways involved in inflammation. These drugs have been shown to reduce disease severity and delay disease progression significantly, particularly in moderate to severe cases [6, 7]. However, the high cost and potential adverse effects, including an increased risk of infections and malignancies, limit their widespread use, particularly in low-income settings like Pakistan.

In addition to pharmacological treatment, surgical interventions remain necessary in cases where medical therapy fails. Approximately 50% of CD patients require at least one surgical procedure within 10 years of diagnosis due to complications such as intestinal obstruction or perforation.

Advances in minimally invasive laparoscopic techniques have improved post-surgical outcomes, but surgery does not cure the disease, as inflammation often recurs in other areas of the intestine. Emerging treatment strategies, including stem cell therapy and microbiome-based interventions such as fecal microbiota transplantation (FMT), offer promising alternatives for disease management, though more clinical trials are needed to establish their efficacy and safety [8, 9].

The exact cause of Crohn's disease remains unknown, but research suggests that it results from a complex interplay of genetic susceptibility, immune dysfunction, environmental factors, and alterations in gut microbiota. More than 200 genetic loci have been associated with CD, with mutations in the NOD2 gene being the most significant genetic risk factor. NOD2 mutations impair bacterial sensing and immune regulation, leading to an exaggerated immune response against commensal gut microbes.

Other key genetic variants implicated in CD include ATG16L1 and IL23R, which are involved in autophagy and immune signaling pathways, respectively [10, 11]. However, genetics alone cannot explain the rising incidence of CD, particularly in regions where the disease was historically rare, indicating that environmental factors such as smoking, diet, stress, and antibiotic use play significant roles in disease onset and progression.

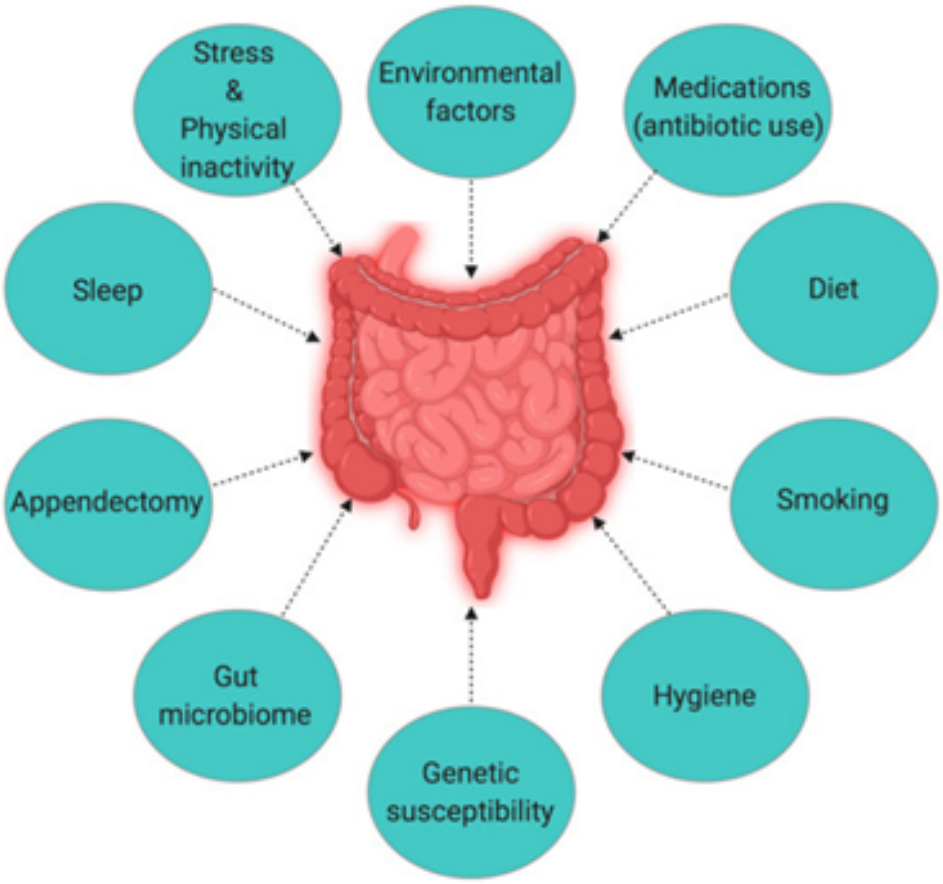

FIGURE 1.2: Genetic and Non genetic Risk Factors associated with Crohn's Disease

Smoking is one of the strongest environmental risk factors for CD, with smokers being twice as likely to develop the disease compared to non-smokers. Additionally, a Westernized diet high in refined sugars, saturated fats, and food additives has been linked to an increased risk of CD by disrupting gut barrier function and promoting pro-inflammatory bacterial growth. Furthermore, the overuse of antibiotics, particularly during childhood, has been associated with an increased risk of developing CD later in life, likely due to its impact on gut microbial composition [12, 13]. Psychological stress has also been identified as a potential trigger for disease flare-ups, suggesting a link between the brain-gut axis and CD pathogenesis. Understanding these environmental triggers is crucial for developing preventive strategies and early interventions [14].

The gut microbiota, composed of trillions of bacteria, viruses, and fungi, plays a

vital role in digestion, immune modulation, and maintaining intestinal homeostasis. In healthy individuals, a balance between beneficial and harmful microbes is maintained, but in CD patients, gut dysbiosis occurs, characterized by a significant reduction in anti-inflammatory bacteria such as *Faecalibacterium prausnitzii* and *Roseburia*, along with an overgrowth of pro-inflammatory bacteria such as *Escherichia coli* and *Fusobacterium nucleatum* [15, 16]. This microbial imbalance leads to increased intestinal permeability, triggering an immune response that exacerbates inflammation and tissue damage.

Recent studies using next-generation sequencing techniques have provided deeper insights into the functional role of the microbiome in CD. Metagenomic analysis has revealed that CD-associated microbiomes exhibit increased production of bacterial endotoxins and reduced short-chain fatty acids (SCFAs), which are essential for gut epithelial health. SCFAs, particularly butyrate, play an anti-inflammatory role by promoting the integrity of the intestinal barrier and regulating immune responses. The loss of butyrate-producing bacteria in CD patients has been linked to increased intestinal inflammation and disease severity [17, 18]. These findings highlight the potential for microbiome-targeted therapies, including probiotics, prebiotics, and microbiota transplantation, as novel treatment strategies for CD.

Metagenomics is a powerful approach that allows for the comprehensive analysis of microbial communities in the gut by sequencing all genetic material present in a sample. Unlike traditional culture-based methods, metagenomics provides an unbiased view of microbial diversity and functional capabilities. Shotgun metagenomics, one of the most widely used techniques, enables the identification of bacterial species, their metabolic pathways, and their interactions with the host. This technique has been instrumental in identifying microbial signatures associated with CD and distinguishing between active disease states and remission [19, 20].

Recent advancements in computational biology have further improved the accuracy of metagenomic analysis, allowing researchers to study microbial gene expression (metatranscriptomics), protein functions (metaproteomics), and metabolite production (metabolomics). These multi-omics approaches provide a more holistic

understanding of the gut microbiome's role in CD and offer potential biomarkers for early diagnosis and treatment monitoring. As metagenomics continues to evolve, integrating it with clinical data and machine learning algorithms may lead to the development of personalized medicine approaches for CD patients, improving disease management and treatment outcomes [21, 22].

1.1 Problem Statement

In Crohn's Disease, the gut microbiota, the population of microorganisms in the digestive tract, is frequently disrupted to create the condition of dysbiosis. The imbalance is one of diminished diversity, low counts of health-promoting bacteria, and higher abundance of potentially pathogenic bacteria. The gut microbiota is central to Crohn's Disease, contributing potentially to chronic inflammation and other disease complications.

1.2 Research Objectives

The study is designed to identify taxonomic biomarkers for Crohn's Disease

To achieve the aim following objectives are designed:

1. To determine the number and taxonomic makeup of the human gut microbiome in individuals with Crohn's Disease
2. To explore taxonomic diversity within Crohn's disease-specific metagenomic samples
3. To perform Multivariable analysis to identify the association between microbial taxa and clinical metadata.

Chapter 2

Literature Review

Inflammatory Bowel Disease, also referred to as IBD, is a group of diseases characterized by swelling and inflammation of the gastrointestinal tract, resulting from an autoimmune disorder in which the immune system attacks healthy bowel cells, particularly in the intestines. Because it is an autoimmune disorder, it is typically a chronic disease. The most common disease conditions under the umbrella of IBD are ulcerative colitis and Crohn's disease.

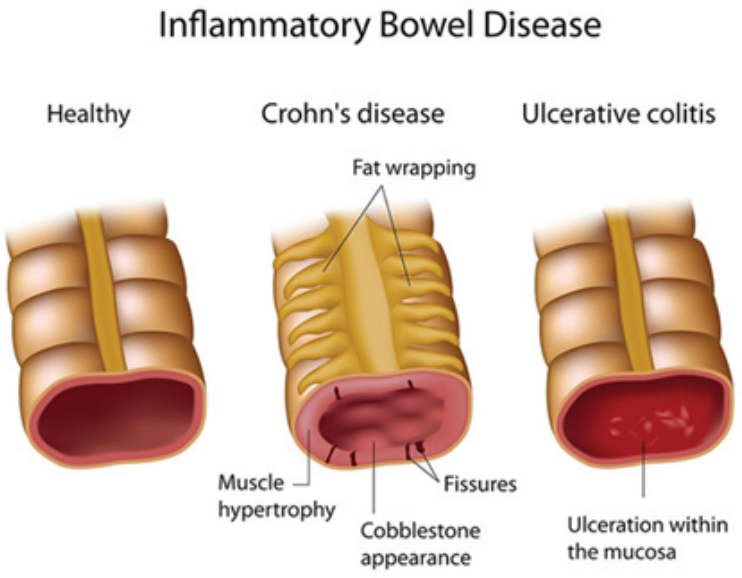

FIGURE 2.1: Inflammation and ulceration conditions of the intestines during Inflammatory Bowel Disease

Although the symptoms, including stomach aches, fatigue, extensive weight loss, and diarrhea, are common in both types of IBD. Severity of these symptoms also varies from very mild disease to fatal complications.

Ulcerative colitis is a type of IBD when ulcers, including inflammation and sores in the colon and rectum while Crohn's disease is a condition where inflammation is in the whole gastrointestinal tract, more frequently in the small intestine than the upper gastrointestinal tract, especially in the deeper layers of epithelium.

TABLE 2.1: Comparison of symptoms and complications in Crohn's Disease and Ulcerative colitis.

Crohn's Disease	**Ulcerative Colitis**
Occurs anywhere along the gastrointestinal tract – from the mouth to the anus; though it most commonly affects the end of the small intestine	Occurs only in the colon and rectum
Affects all layers of the intestinal walls	Only affects the innermost lining of the intestine
Inflamed areas can appear in parts, even next to perfectly healthy parts of the intestines	Inflamed areas are continuous
Possible complications include: bowel obstruction, anal fissures, colon cancer, strictures, fistulas, and malabsorption	Possible complications include: colon perforation, colon cancer, severe dehydration and toxic megacolon

Stress and diet are the major contributors to IBD, but several other factors also contribute to the severity of the disease, including a dysfunctional immune system or autoimmune disease, genetic factors, sensitivity of the digestive system, infections during childhood, excessive use of antibiotics, and poor hygiene, especially for bottle-fed kids.

Therefore, in general, it could be concluded that, in addition to family history and genetics, the extensive use of nonsteroidal anti-inflammatory medicine is the major contributor in IBD.

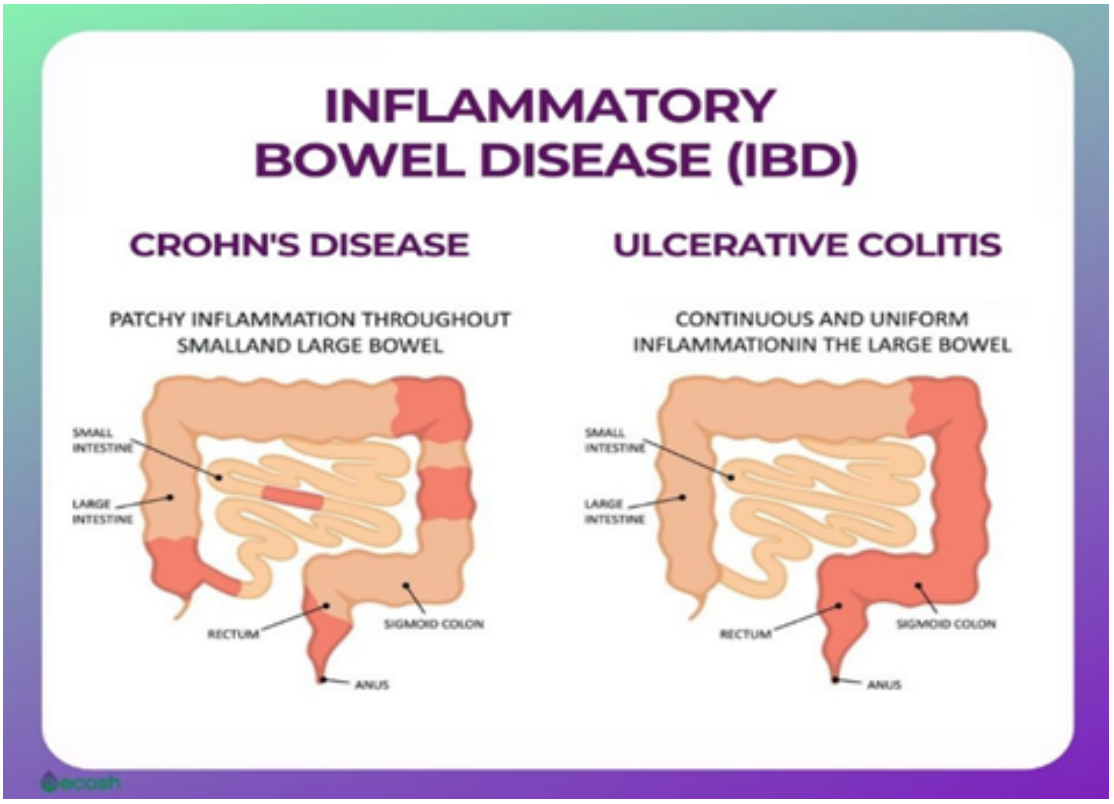

FIGURE 2.2: Comparison of the location of inflammation in Crohn's disease and Ulcerative colitis

Both types of IBD, either ulcerative colitis or Crohn's disease, ultimately result in severe complications, including Colon Cancer, which is the major complication with onset between 8-10 years of IBD. Arthritis and swelling of the skin are also commonly observed in patients with IBD.

Patients diagnosed with ulcerative colitis face toxic megacolon, i.e., swelling and rapid widening of the colon, and perforated colon i.e., holes in the colon. Anal fissures are very common in people suffering with Crohn's disease along with development of fistulas especially around anal regions.

2.1 Crohn's Disease: Definition

Crohn's disease (CD) has been defined by the World Health Organization (WHO) as a chronic, relapsing inflammatory bowel disease (IBD) that may involve any

region of the gastrointestinal (GI) tract, producing inflammation, ulceration, and destruction of the intestinal wall.

The disease is marked by periods of remission and flare-ups, followed by complications like intestinal strictures, fistulas, and disorders of malabsorption [1]. Unlike ulcerative colitis, which is limited to the colon, Crohn's disease can affect any area from the mouth to the anus [2].

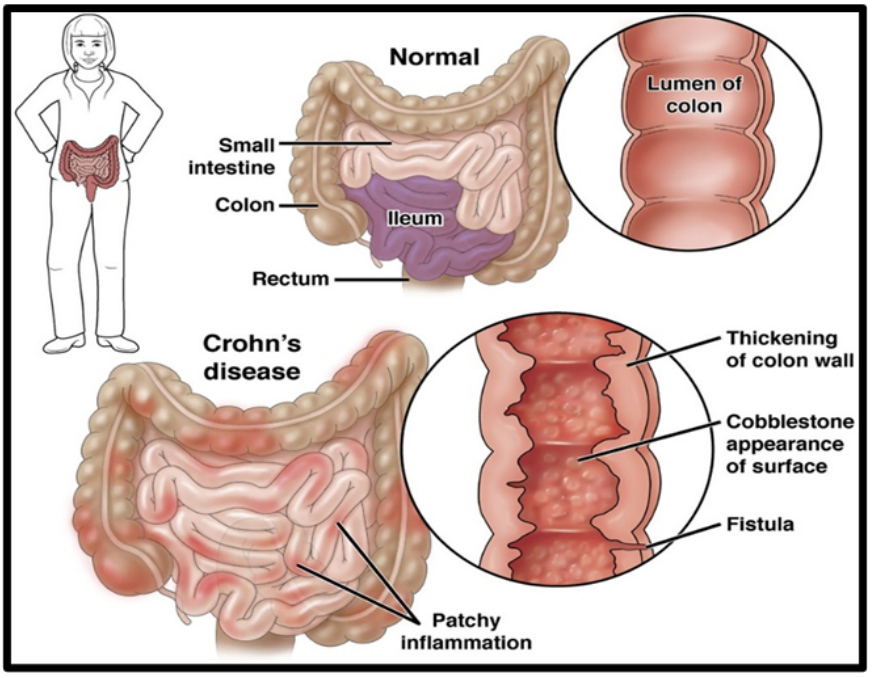

FIGURE 2.3: Location of disease infections [23]

2.2 Prevalence of Crohn's Disease

2.2.1 Global Prevalence

Western countries are more popular for Crohn's Disease, with the highest incidence in North America (319 per 100,000) and Europe (322 per 100,000). However, recent studies indicate a rising incidence in Asia, Africa, and the Middle East, likely due to urbanization and dietary changes [24].

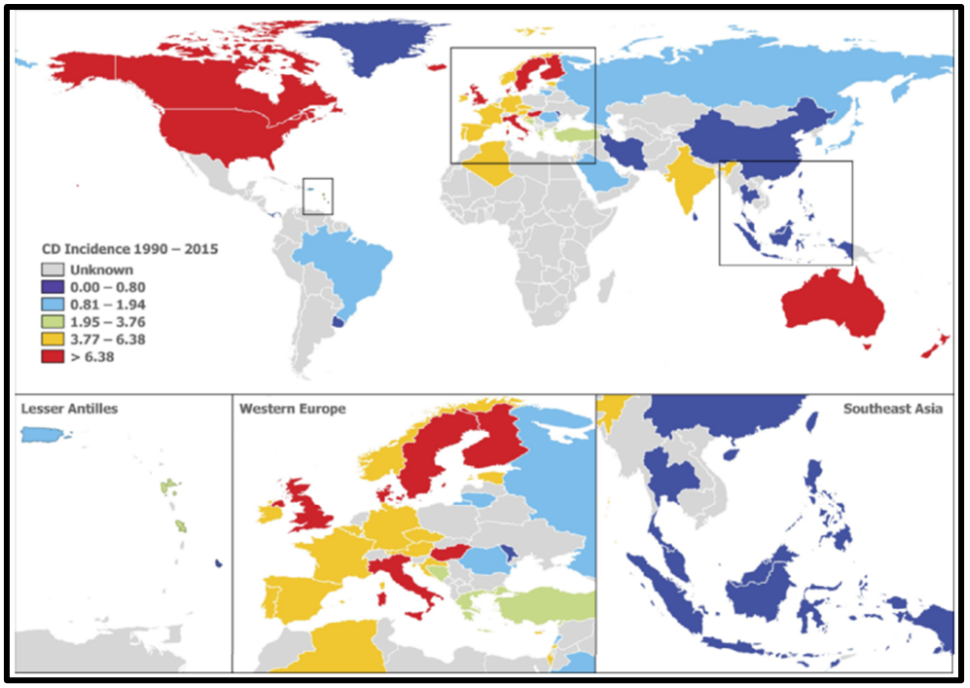

FIGURE 2.4: Worldwide map of incidence of Crohn's disease [25]

2.2.2 Prevalence in Pakistan

In Pakistan, Crohn's disease was historically underreported, but recent studies suggest an increasing incidence, particularly in urban areas. A hospital-based study found that 20% of IBD cases in Pakistan are Crohn's disease [3].

2.2.3 Gender & Age-Wise Prevalence

Crohn's disease can occur at any time in life, but it is most frequently diagnosed between 15 and 35 years of age. Both sexes are equally affected by the disease, although according to some reports, there is a minor male pre-dominance in younger age groups and a higher female prevalence after age 50 [4].

2.3 Diagnostics for Crohn's Disease (CD)

The examination of Crohn's disease involves a combination of clinical symptoms, imaging, endoscopy, and laboratory tests. The Fecal Calprotectin Test serves as

a biomarker that indicates intestinal inflammation. Colonoscopy with biopsy is considered the gold standard for confirming Crohn's disease, as it allows for the identification of ulcers, strictures, and granulomas in the gastrointestinal tract. Magnetic Resonance Enterography (MRE) is one of the imaging methods for visualizing strictures and inflammation, especially in the small intestine. Serology also plays an important role in differentiating between ulcerative colitis and Crohn's disease by identifying certain antibody markers like ASCA (anti-*Saccharomyces cerevisiae* antibodies) and pANCA (perinuclear anti-neutrophil cytoplasmic antibodies) [5].

TABLE 2.2: Clinical Presentation of Crohn's Disease by Affected Site

Location Affected	**Common Symptoms**	**Clinical Notes**	**Prevalence (%)**
Ileum and Colon	Diarrhea, cramping, abdominal discomfort, unintentional weight loss	Most frequently observed form	35%
Colon Alone	Diarrhea with rectal bleeding, perirectal pain, fistulae, anal ulcers	More prone to skip lesions and joint-related symptoms	32%
Small Intestine Only	Cramping, abdominal pain, diarrhea, weight loss	Can lead to serious complications such as abscesses or fistulas	28%
Gastroduodenal Region	Loss of appetite, nausea, vomiting, weight loss	Least common type; may result in bowel obstruction	5%

2.4 Available Treatments for Crohn's Disease

2.4.1 Pharmacological Treatments

Aminosalicylates are commonly used for mild cases of Crohn's disease (CD), but they are less effective compared to their use in ulcerative colitis. Corticosteroids

are very effective in decreasing inflammation, but when used in the long term, they may have some important side effects.

Immunosuppressants, such as Azathioprine and Methotrexate, help control Crohn's disease by reducing immune system overactivation. Biologic therapies, including Infliximab and Adalimumab, specifically target the tumor necrosis factor-alpha (TNF-α) for effective modulation of inflammation and, subsequently, disease progression [6].

2.4.2 Surgical Interventions

Approximately 50% of CD patients require surgery due to complications such as intestinal strictures or fistulas. The most common procedure is ileocecal resection, where the diseased portion of the intestine is removed [7].

2.4.3 Emerging Therapies

Fecal Microbiota Transplantation (FMT) treatment technique that seeks to re-establish the balance of gut microbiota through the transfer of healthy donor fecal microbiota into the gut of Crohn's disease patients.

Stem cell therapy is currently under investigation as a possible treatment of Crohn's disease, with a focus on its role in regulating immune responses and promoting intestinal tissue repair.

2.5 Causes of Crohn's Disease

2.5.1 Genetic Causes

Over 200 genetic loci are linked with the disease called Crohn's disease or CD. The NOD2 gene mutation is the most significant genetic risk factor, impairing

bacterial recognition and immune response. Other genes involved in CD include IL23R (inflammatory response) and ATG16L1 (autophagy regulation) [8].

2.5.2 Non-Genetic Causes

Smoking is a significant threat to Crohn's disease, increasing its likelihood of developing the condition by two to three times. Dietary habits also play a crucial role, with high-fat and processed diets contributing to increased gut inflammation and worsening disease symptoms.

Additionally, gut dysbiosis, linked with a reduction in *Faecalibacterium prausnitzii*, which is good bacteria, results in heightened inflammation and disruption of intestinal homeostasis [9].

2.6 Risk Factors of Crohn's Disease

Crohn's disease is a major threat to patients with a family history of this disease, with first-degree relatives of affected individuals having a 5 to 10 times higher risk of developing the condition.

The application of antibiotics, especially in childhood, has been found to increase the risk of Crohn's disease, likely due to its impact on gut microbial composition.

An imbalance in gut microbes is characterized by a rise in Proteobacteria and a fall in Firmicutes, disrupts gut homeostasis, and contributes to chronic intestinal inflammation [10].

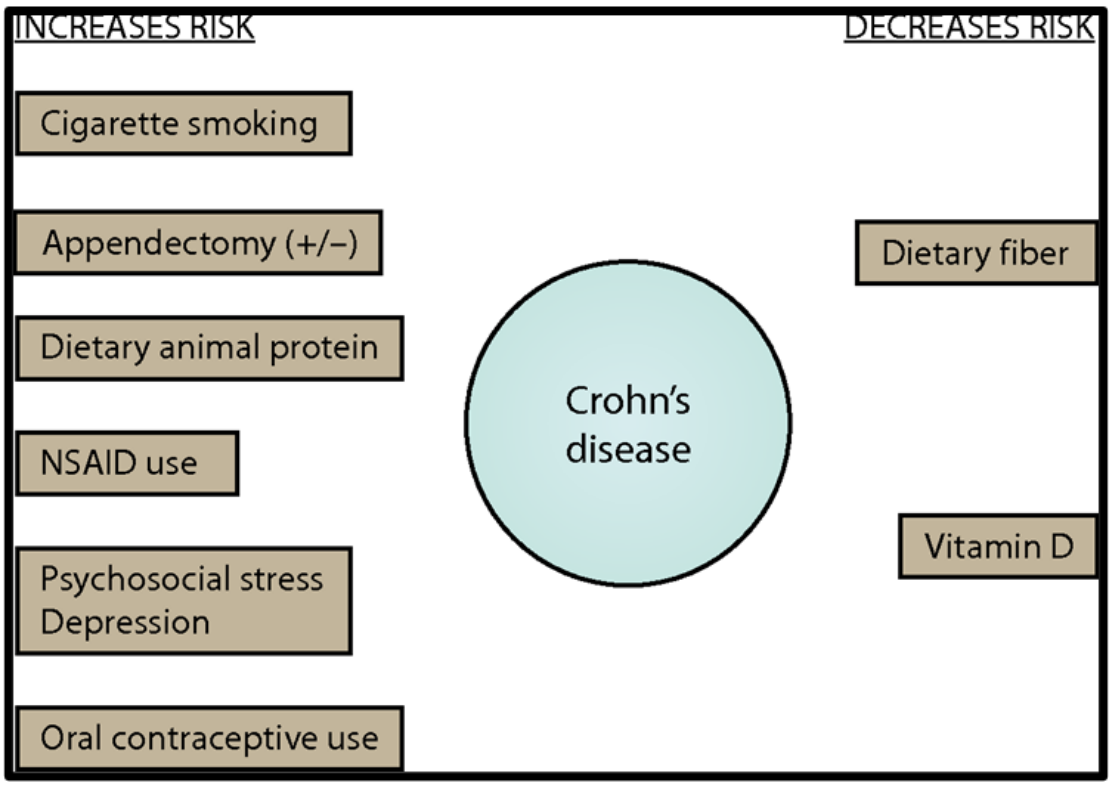

FIGURE 2.5: Risk factors of Crohn's Disease [27]

2.7 Gut Microbiota's Role in Crohn's Disease

The gut microbiota has a central role in CD pathogenesis:

Dysbiosis, an imbalance of microbes in the gut, is a central characteristic of Crohn's disease and is defined by the loss of protective bacteria like *Faecalibacterium prausnitzii* and overgrowth of disease-causing bacteria like Escherichia coli. Immune modulation is strongly regulated by metabolites of microbes, most notably short-chain fatty acids (SCFAs), which are important in the regulation of inflammation and the maintenance of gut homeostasis.

Moreover, increased permeability of intestines, also termed "leaky gut," ensues due to dysbiosis, resulting in increased immune activation and chronic inflammation in Crohn's disease patients [11].

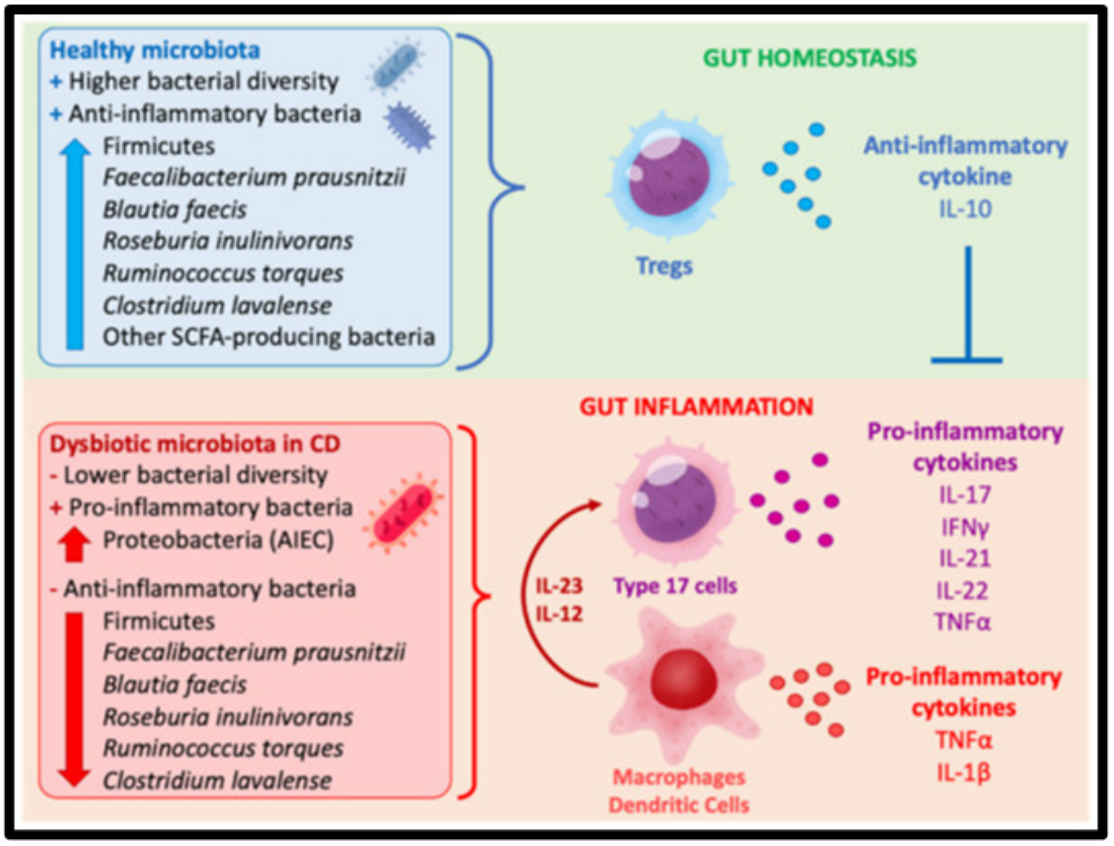

FIGURE 2.6: An overview of the inflammatory mechanisms contributing to Crohn's disease progression with gut dysbiosis [28]

2.8 Metagenomics Analysis in Crohn's Disease

To analyze the role of bacterial species or infection in the disease, a traditional genomics approach is utilized, the exploration of complete genetic information of a single selected microbial species, while to have a complete picture of all the microbial species involved in the disease, a metagenomic approach is used. The metagenomic approach utilizes two approaches, including sequence-based analysis and functional-based analysis.

The comprehensive information provided by metagenomic analysis includes alpha diversity which is diversity of species within single sample i.e. in this case the microbial species present in a patient suffering with Crohn's disease. Beta diversity on other hand provides the difference in microbial community composition

between different samples i.e. the differences of microbial species between healthy individuals and Crohn's disease patients.

TABLE 2.3: Metagenomics enables culture-independent microbiome analysis

Step	Description
DNA Extraction	Collecting microbial DNA from stool samples
Shotgun Sequencing	High-throughput sequencing of entire microbial genomes
Bioinformatics Analysis	Identifying bacteria and their functional pathways
Statistical Modeling	Correlating microbiome changes with CD severity [12]

2.9 Taxonomic Classification of Bacteria in CD

As gut microbial species are among the major contributors in the risks of Crohn's disease, the metagenomic profiling can help us understand the role that how these microbial species are contributing in disease onset and severity of symptoms. The published studies reveal that in the case of Crohn's disease not only the bacterial species but virome also plays a significant role. But more reliabel data is available in case of bacterial species where they contribute functional in bacterial host interactions, carbohydrate metabolis and alterations in many pathways.

TABLE 2.4: Metagenomic studies classify bacteria using 16S rRNA sequencing

Phylum	Change in CD Patients	Effect
Firmicutes	↓ Decreased	Lower SCFA production
Proteobacteria	↑ Increased	Gut inflammation
Bacteroidetes	↓ Decreased	Impaired digestion
Actinobacteria	↑ Increased	Pathogenic overgrowth [13]

2.10 Multivariate Association Studies in CD

Multivariate analysis is a powerful statistical approach used to identify correlations between the microbiome and Crohn's disease. PERMANOVA (Permutational Multivariate Analysis of Variance) is often used to test changes in microbial diversity in healthy subjects and patients with Crohn's disease. LEfSe (Linear

Discriminant Analysis Effect Size) is another key analytical method that helps identify specific bacterial biomarkers associated with disease progression. Additionally, machine learning or ML models are increasingly being utilized to forecast disease outcomes by examining complex microbiome reads and identifying patterns linked to Crohn's disease severity and treatment efficacy [14].

Chapter 3

Research Methodology

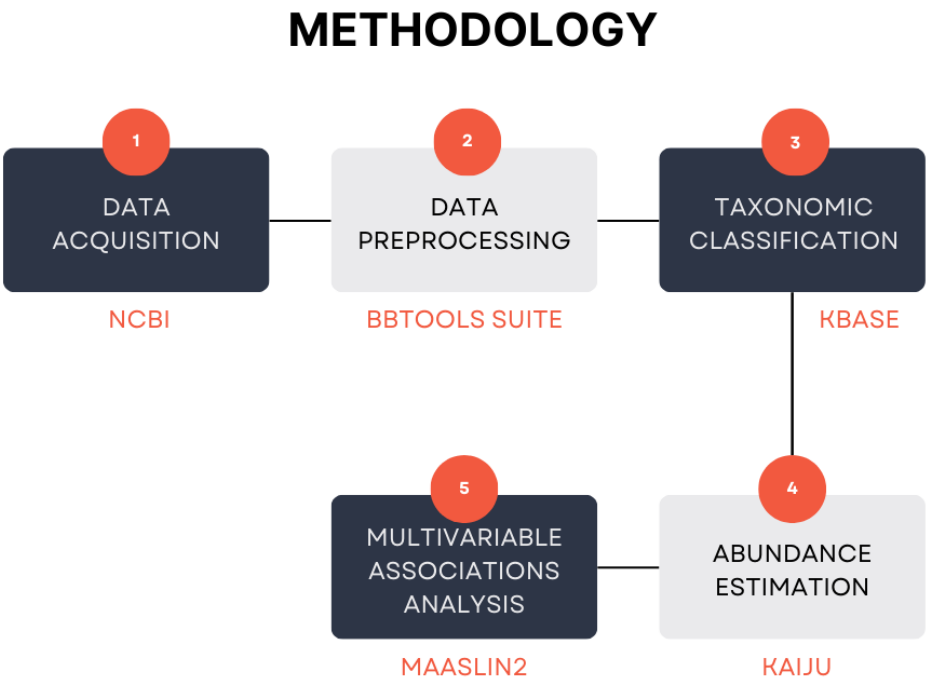

FIGURE 3.1: Research Methodology

3.1 Data Acquisition

To investigate taxonomic biomarkers and microbial associations in Crohn's Disease (CD), publicly available Whole Genome Shotgun (WGS) sequencing data

were retrieved from the NCBI Sequence Read Archive (SRA) under Project ID PRJNA400072. This dataset included, To support downstream multivariable association analyses, corresponding metadata were obtained from the NCBI Bio Sample database. These metadata included relevant clinical and demographic information necessary for statistical modeling and association studies. All raw sequencing data were retrieved and processed using the KBase online platform. The "*Import SRA File as Reads from Web - v1.0.10*" tool was employed to efficiently import sequencing reads directly from the SRA repository.

3.2 Data Preprocessing

Data preprocessing is a critical step in the data analysis pipeline, ensuring data quality and integrity for accurate and meaningful analysis. Proper preprocessing enhances the reliability of downstream computational analyses, leading to more robust and reproducible results. Ensuring high-quality sequencing data is crucial for reliable downstream analyses. The preprocessing steps included quality control (QC) of sequencing reads and removal of host-derived sequences to minimize contamination.

3.3 Quality Control (QC)

Quality control was performed on all samples using FastQC (v0.12.1) to evaluate key sequencing metrics, including:

- Per-base sequence quality scores
- GC content distribution
- Adapter contamination
- Overall read statistics (read length, duplication levels, etc.)

This step ensured that low-quality reads and potential sequencing artifacts were identified before further processing.

3.4 Adapter and Low-Quality Base Removal

To eliminate adapter sequences and trim low-quality bases, Trimmomatic (v0.39) was employed with the following parameters:

Adapter Clipping Parameters

- Adapter type: TruSeq2-PE
- Seed mismatches: 2 (Maximum mismatch count for aligning adapter sequences)
- Palindrome clip threshold: 30 (Threshold for removing adapter fragments in paired end reads)
- Simple clip threshold: 10 (Threshold for removing adapter sequences in single end reads)

3.5 Host Read Removal

Metagenomic datasets often contain contaminant sequences originating from the host organism, which can lead to false-positive results in microbial analysis. Since this study focuses on the gut microbiome, human-derived reads were removed to ensure that only microbial sequences were retained. The BBDuk script (included in the BBTools suite) was used to filter out host-derived reads by mapping sequencing reads against the GRCh38 human genome reference. The steps involved:

- Aligning metagenomic reads to the GRCh38 reference genome.
- Removing host-aligned reads from the dataset.

- Retaining only non-host microbial reads for subsequent analysis.

After host read removal, FastQC was rerun to reassess the quality of the post-processed data, ensuring that the dataset was clean and suitable for taxonomic profiling and multivariable association analysis.

3.6 Taxonomic Classification

Taxonomic classification of metagenomic reads was performed using Kaiju v1.9.0 on the KBase online platform, a tool renowned for its protein-level classification approach, which significantly enhances sensitivity in detecting microbial taxa from complex metagenomic datasets. For this analysis, Kaiju's default reference database, comprising either the NR or RefSeq database, was employed to ensure comprehensive and accurate taxonomic identification. To balance precision and computational efficiency, a minimum match length of 11 amino acids was applied, ensuring that only sequences meeting this threshold were classified. Additionally, to optimize processing time, 10% of the total reads were randomly subsampled for analysis, with a single replicate run (subsample replicates set to 1) to maintain consistency in the subsampling process. Taxonomic assignments were generated across six hierarchical levels: Phylum, Class, Order, Family, Genus, and Species, providing a detailed and structured representation of the microbial community composition within the dataset. The use of Kaiju v1.9.0 ensured robust and efficient taxonomic profiling while maintaining high sensitivity in identifying diverse microbial taxa.

3.7 Abundance Estimation

Kaiju performed abundance estimation to determine the relative distribution of microbial taxa within the dataset. The number of reads assigned to each taxon was recorded across six taxonomic levels: Phylum, Class, Order, Family, Genus, and Species, providing a comprehensive overview of microbial composition. Relative

abundance was calculated by determining the proportion of reads assigned to each taxon relative to the total number of classified reads. To minimize noise and focus on biologically relevant taxa, a filtering threshold of 0.5% relative abundance was applied. Taxa with less than 0.5% relative abundance were excluded from the final output, ensuring that only the most significant and abundant microbial groups were retained for downstream analysis. This approach allowed for a clearer and more accurate representation of the microbial community structure within the dataset.

3.8 Multivariable Associations Analysis

To identify associations between microbial taxa and key clinical metadata, including age, diagnosis (CD vs. Healthy), mesalamine usage, and immunosuppressant treatment, a multivariable association analysis was performed using MaAsLin2. This analysis employed a linear regression model to examine the relationship between microbial species' relative abundance and clinical variables while adjusting for potential confounders.

Each microbial species served as a dependent variable, and clinical metadata were used as independent variables in the model. The regression coefficients (Coef) represented the direction and magnitude of the association, where positive coefficients indicated increased abundance in the presence of a given clinical condition (e.g., higher abundance in CD patients), and negative coefficients indicated decreased abundance (e.g., depletion in CD patients).To control for multiple hypothesis testing, False Discovery Rate (FDR) correction was applied using the Benjamini-Hochberg method, reducing the likelihood of false-positive associations. The adjusted q-value was used to determine statistical significance, with $q < 0.1$ set as the threshold for significant associations. To account for potential confounding factors, the model included adjustments for sample size ($N = 10$) and the presence of features (N.not.0), ensuring that only microbial taxa present in a sufficient number of samples were considered. Additionally, normalization was applied where necessary to standardize the data distribution.

Chapter 4

Results and Discussion

4.1 Data Acquisition

This table provides a structured overview of the Whole Genome Shotgun (WGS) sequencing samples used for taxonomic biomarker identification and multivariable association analysis. The dataset consists of five Crohn's Disease (CD) samples and five Healthy Control (HC) samples, retrieved from the NCBI Sequence Read Archive (SRA) under Project ID PRJNA400072.

Table 4.1: Samples of CD and Healthy

Crohn's Disease (CD) Samples	**Healthy Control (HC) Samples**
SRR6468520	SRR6468521
SRR6468527	SRR6468642
SRR6468559	SRR6468646
SRR6468560	SRR6468649
SRR6468561	SRR6468680

4.2 Data Preprocessing

To eliminate adapter sequences and trim low-quality bases, Trimmomatic (v0.39) was used with TruSeq2-PE adapters, allowing up to 2 seed mismatches, a palindrome clip threshold of 30 for paired-end reads, and a simple clip threshold of

10 for single-end reads. After trimming, human-derived reads were removed using BBDuk (BBTools suite) by mapping sequences against the GRCh38 human genome reference, ensuring only microbial reads were retained.

The process involved aligning reads to the reference genome, discarding host-aligned sequences, and keeping non-host microbial reads for further analysis. FastQC was rerun post-processing to confirm data quality

The BBMap command optimizes host read removal by balancing speed and accuracy. The GRCh38 reference genome is used to identify and filter out human-derived reads, with the nodisk option ensuring that processing occurs in RAM for faster execution. A minimum identity threshold (minid=0.95) removes reads with ≥95% similarity to the human genome, while maxindel=3 limits insertions/deletions to 3 bp, improving alignment precision.

Further optimizations include bwr=0.16 for efficient memory allocation and bw=12 to control bandwidth alignment, balancing speed and sensitivity. The quick match and fast options enhance alignment speed without compromising accuracy.

Finally, maxsites=1 ensures each read is aligned to only one location, preventing ambiguous mappings. Together, these parameters enable efficient host read removal, retaining only high-confidence microbial reads for downstream taxonomic and biomarker analysis.

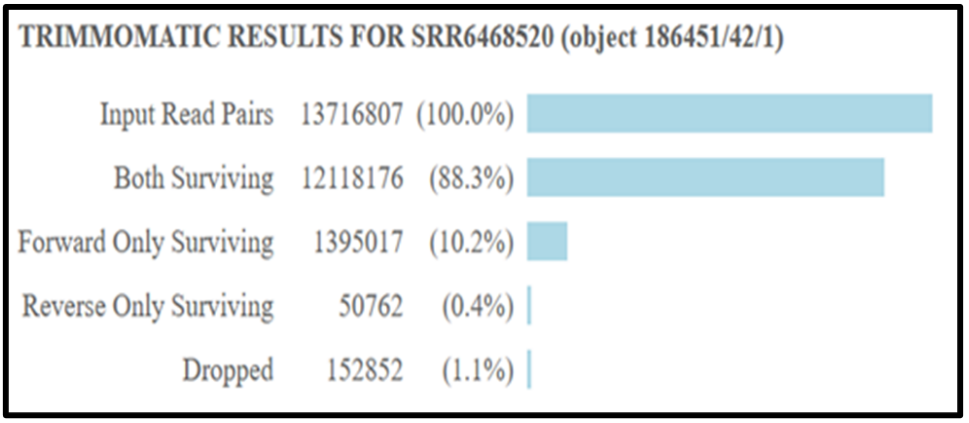

FIGURE 4.1: Trimmomatic Results of Crohn's Disease Sample 1

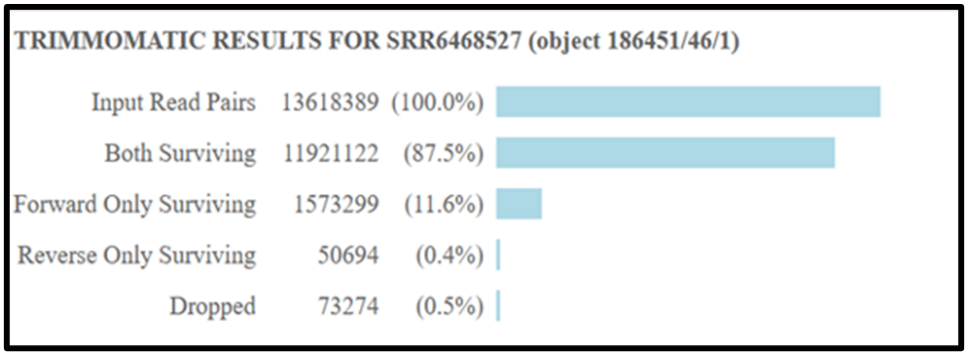

FIGURE 4.2: Trimmomatic Results of Crohn's Disease Sample 2

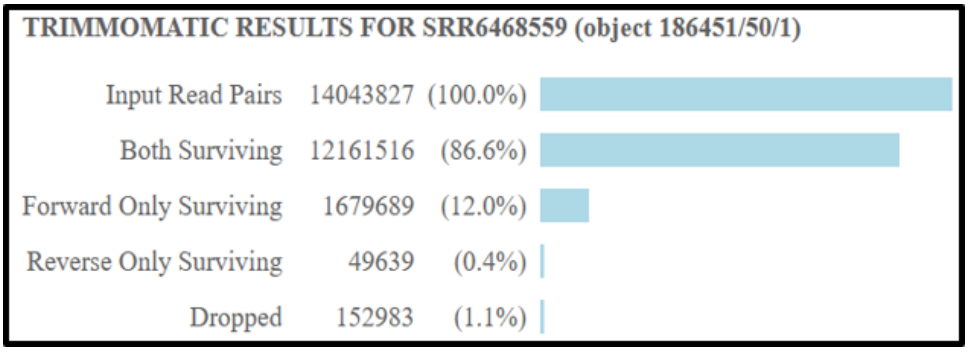

FIGURE 4.3: Trimmomatic Results of Crohn's Disease Sample 3

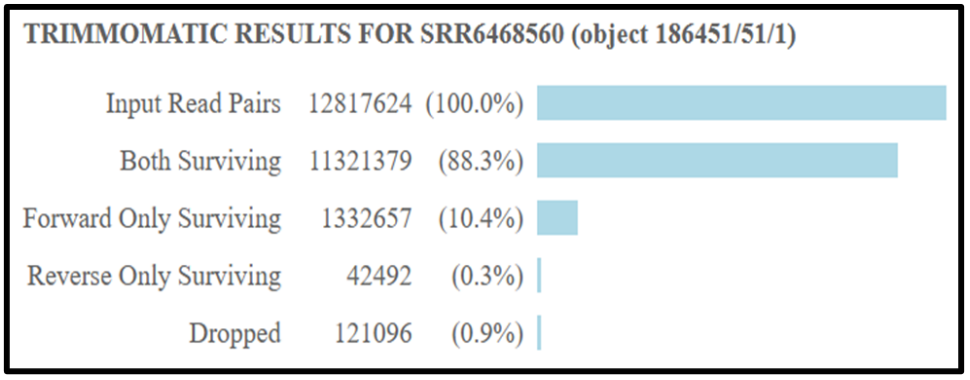

FIGURE 4.4: Trimmomatic Results of Crohn's Disease Sample 4

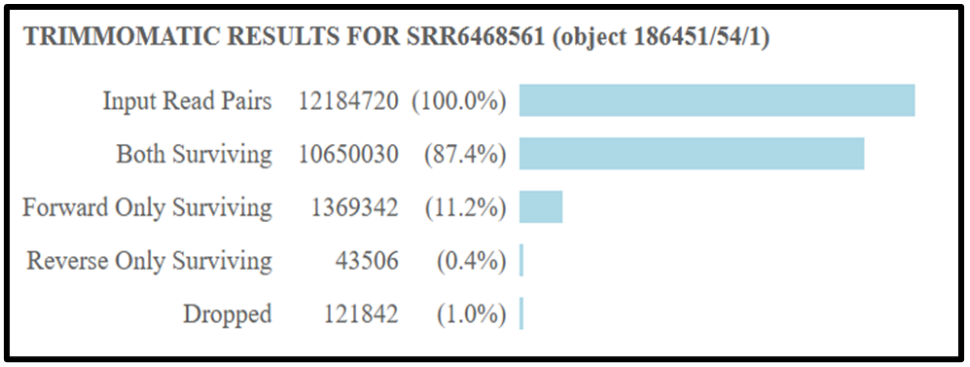

FIGURE 4.5: Trimmomatic Results of Crohn's Disease Sample 5

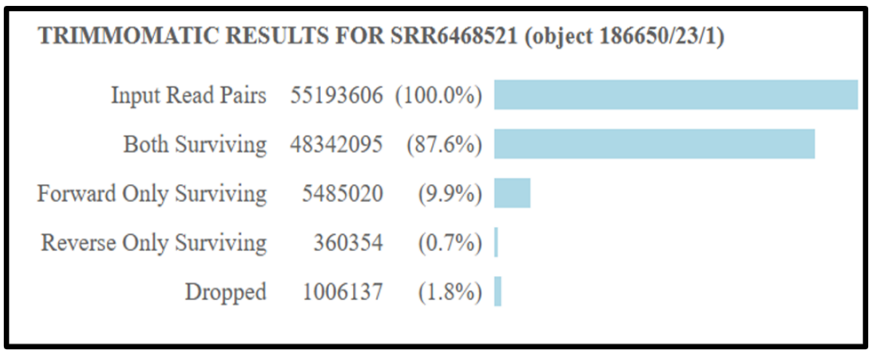

FIGURE 4.6: Trimmomatic Results for Healthy Control Sample 1

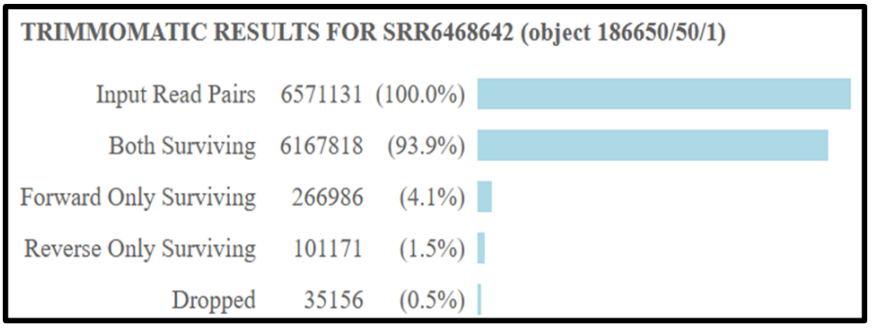

FIGURE 4.7: Trimmomatic Results for Healthy Control Sample 2

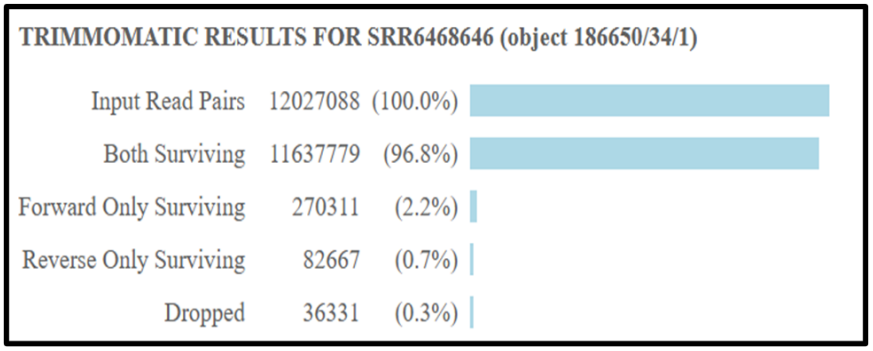

FIGURE 4.8: Trimmomatic Results for Healthy Control Sample 3

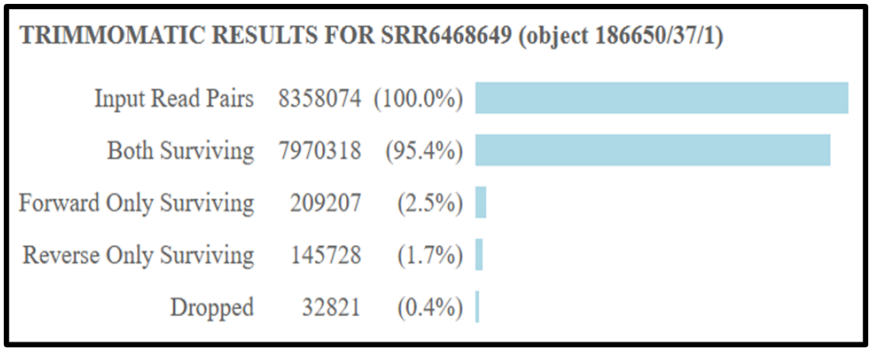

FIGURE 4.9: Trimmomatic Results for Healthy Control Sample 4

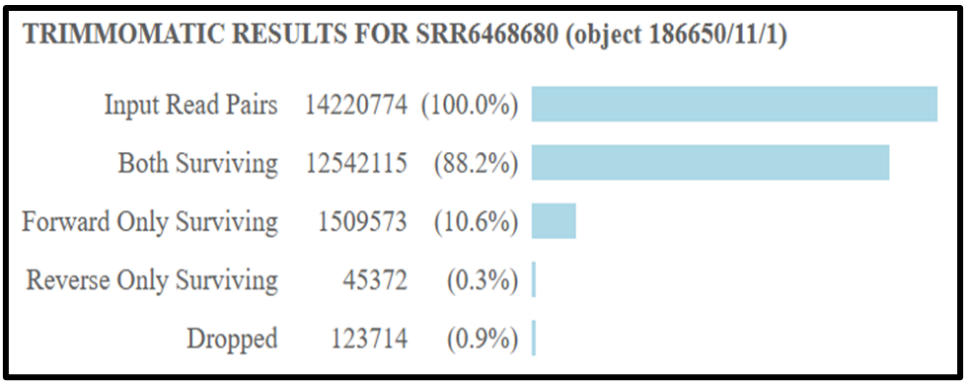

FIGURE 4.10: Trimmomatic Results for Healthy Control Sample 5

4.3 Taxonomic Classification & Abundance Estimation

In comparative analysis of microbial composition between Crohn's Disease (CD) and healthy samples, several key trends emerge. *Faecalibacterium prausnitzii*, a well-known beneficial gut bacterium, is significantly more abundant in healthy samples (17.22%) compared to CD samples (10.03%), highlighting its reduction in CD and its potential role in gut health. Similarly, *Bifidobacterium longum* and *Bacteroides uniformis* show a higher presence in healthy individuals, indicating a decrease in CD samples.

On the other hand, *Anaerostipes hadrus* and *Roseburia intestinalis* exhibit an increase in CD samples, suggesting a possible shift in microbial composition associated with disease progression. Notably, *Lactobacillus gasseri* and *Enterococcus faecium* are detected exclusively in healthy samples and are completely absent in CD samples, which may imply their protective or stabilizing role in gut health. Additionally, *Akkermansia muciniphila* and *Bacteroides stercoris* show a marked decrease in CD samples compared to healthy ones, reinforcing their potential significance in maintaining gut microbial balance.

These variations in microbial abundance suggest that CD is associated with a disruption in the gut microbiome, with a decrease in beneficial bacteria and an increase in certain taxa that may contribute to disease progression.

TABLE 4.2: Comparison of Relative Abundance Between CD and Healthy

Taxon Name	**Relative Abundance in CD (%)**	**Relative Abundance in Healthy (%)**	**Increase in CD (%)**
Faecalibacterium prausnitzii	10.03	17.22	-7.19
Bifidobacterium longum	1.73	4.46	-2.73
Bacteroides uniformis	1.56	3.48	-1.92
Phocaeicola vulgatus	2.31	1.95	0.36
Coprococcus comes	1.8	2.17	-0.37
Anaerobutyricum hallii	1.16	1.6	-0.44
Anaerostipes hadrus	2.43	1.46	0.97
Roseburia intestinalis	2.68	1.12	1.56
Ruminococcus gnavus	3.24	2.79	0.45
Akkermansia muciniphila	2.74	7.33	-4.59
Bacteroides stercoris	2.58	5.4	-2.82
Collinsella aerofaciens	3.23	4.62	-1.39
Bifidobacterium adolescentis	4.31	2.89	1.42
Lactobacillus gasseri	17.68	Not detected	17.68
Enterococcus faecium	9.37	Not detected	9.37

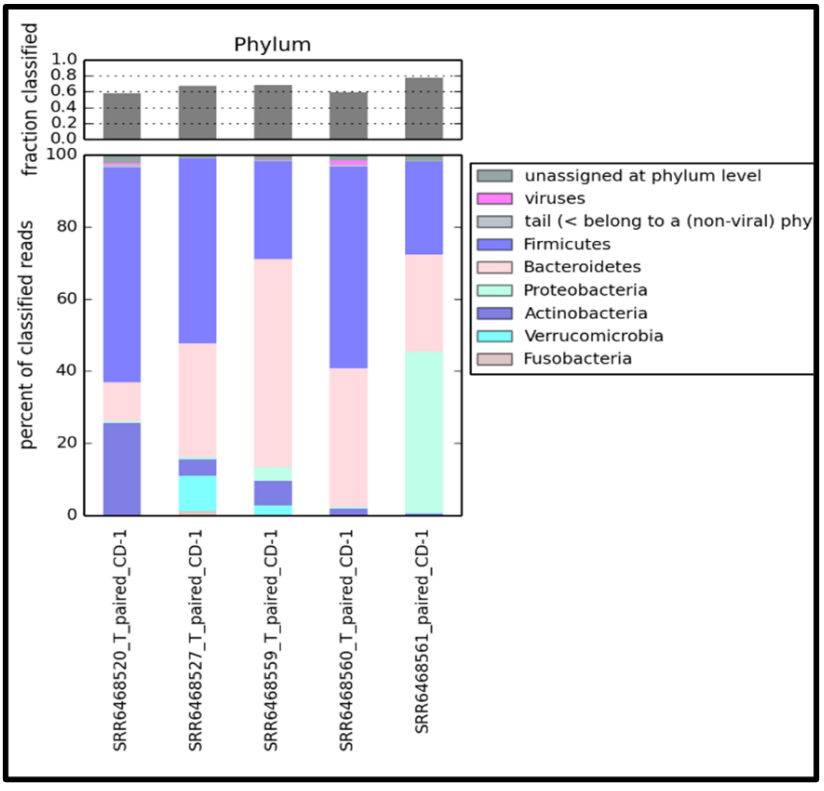

FIGURE 4.11: Classification of Crohn's Disease Sample at Phylum level

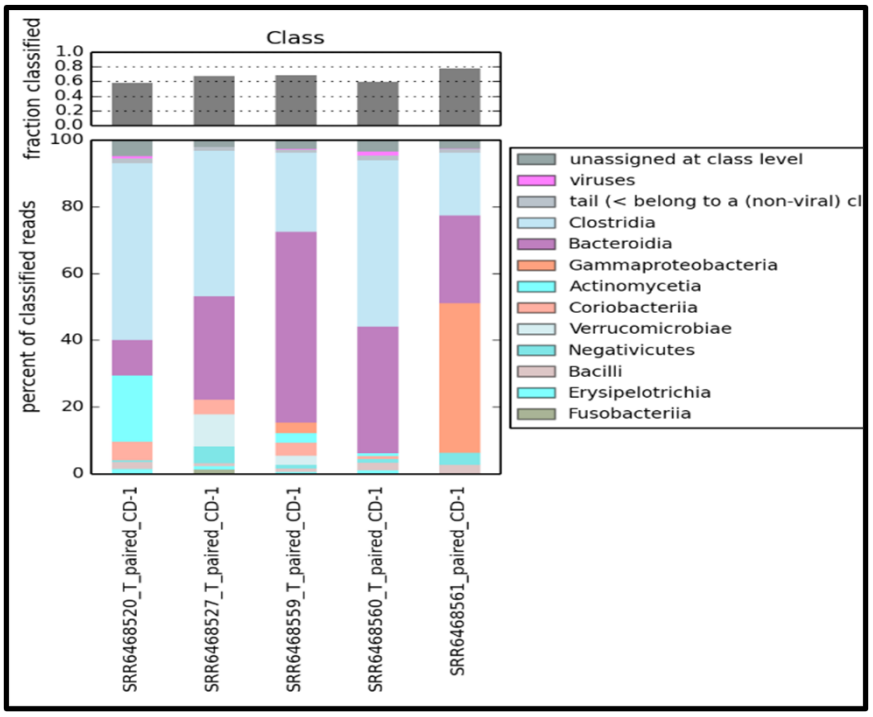

FIGURE 4.12: Classification of Crohn's Disease Sample at Class level

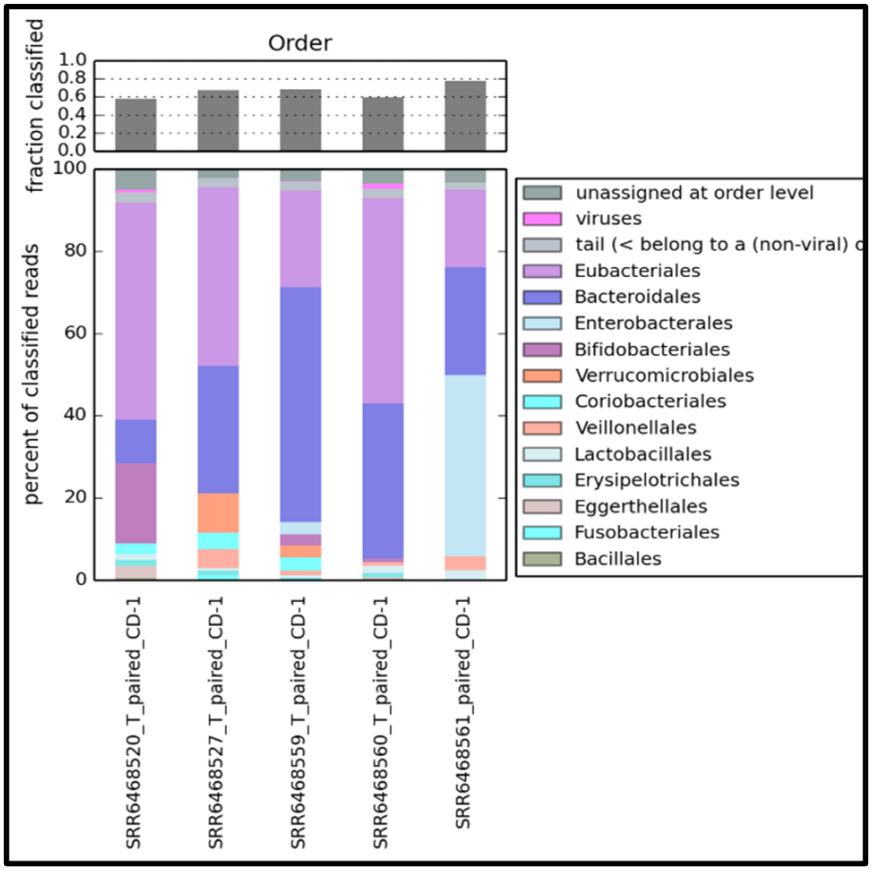

FIGURE 4.13: Classification of Crohn's Disease Sample at Order level

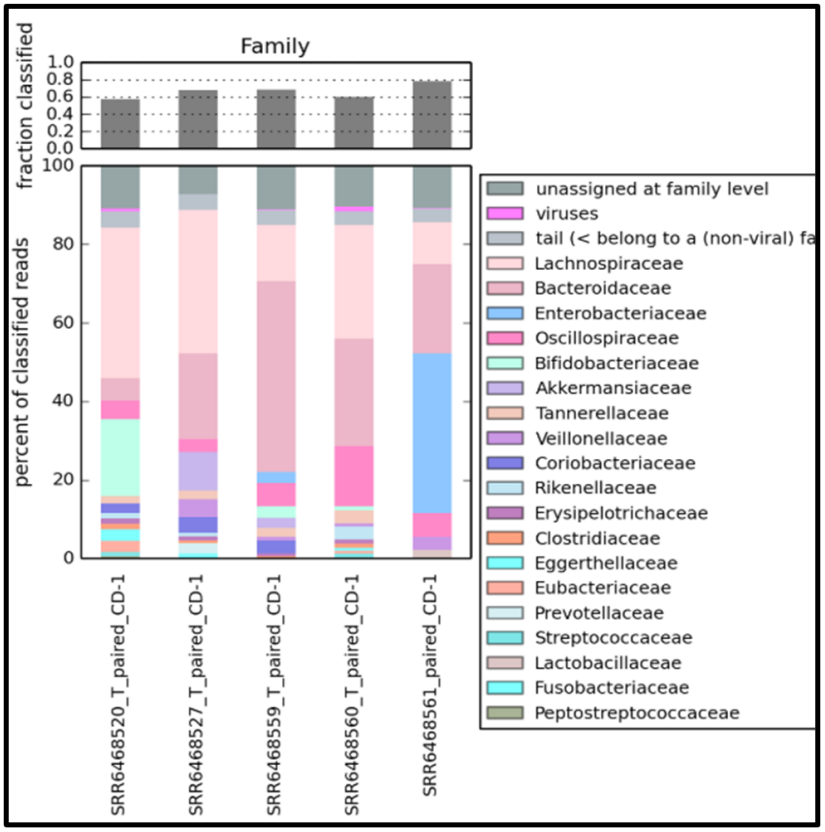

FIGURE 4.14: Classification of Crohn's Disease Sample at Family level

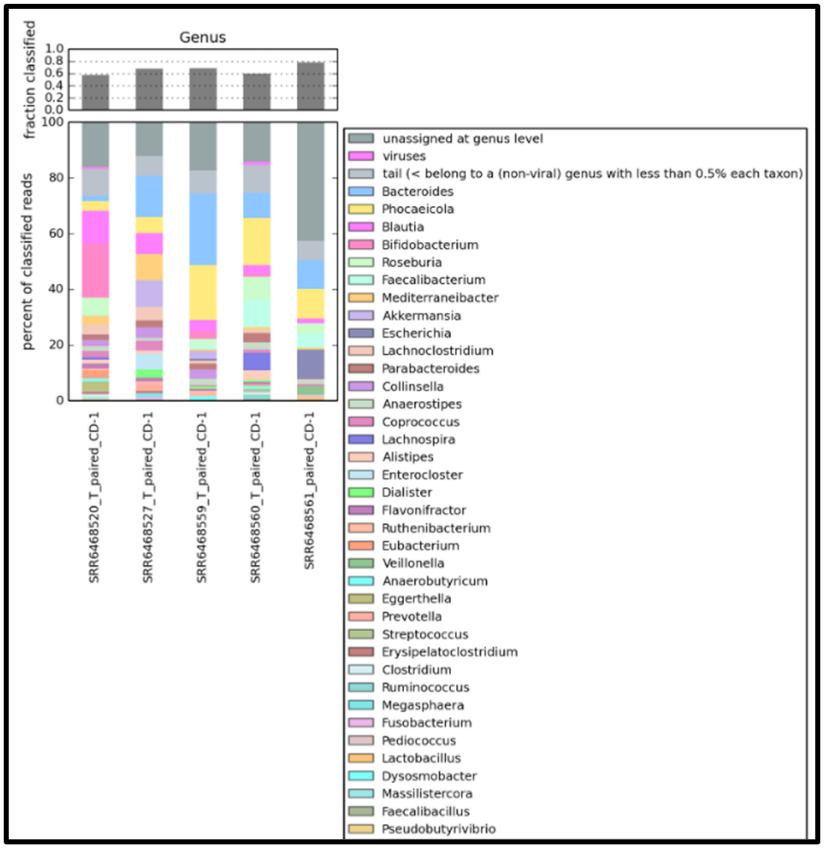

FIGURE 4.15: Classification of Crohn's Disease Sample Genus Level

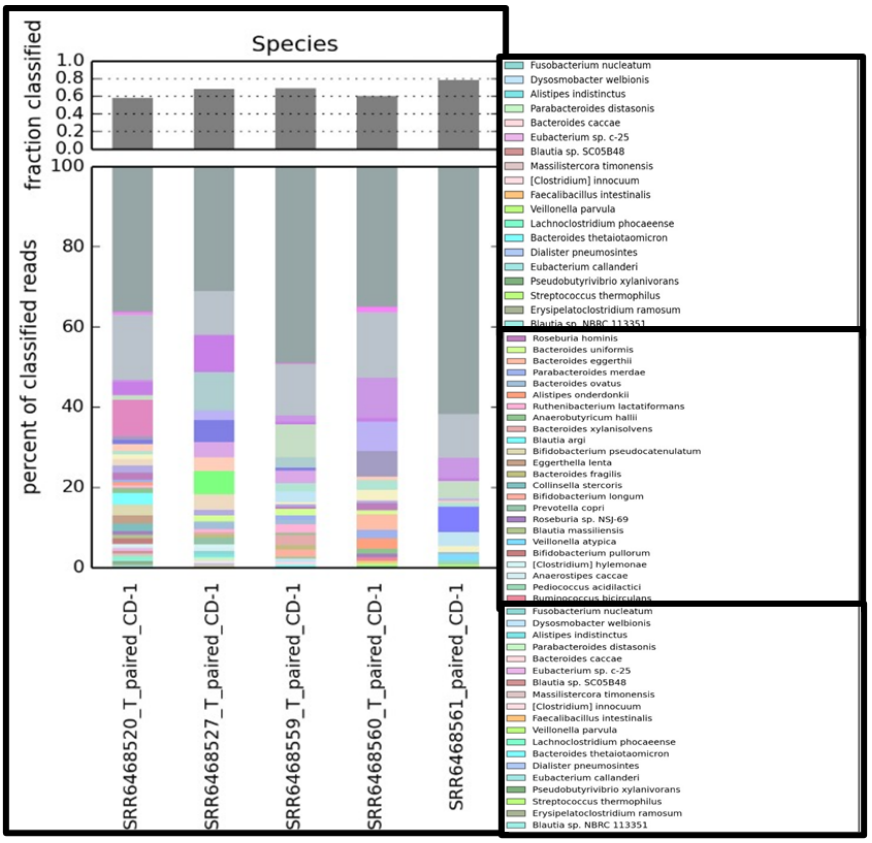

FIGURE 4.16: Classification of Crohn's Disease Sample at Species level

At the phylum level, the gut microbiome was dominated by Proteobacteria and Firmicutes, collectively comprising over 70% of classified sequences, aligning with dysbiosis patterns in Crohn's Disease. Actinobacteria and Bacteroidetes were present in lower proportions (<10%).

At the class level, Gammaproteobacteria (e.g., Enterobacteriaceae) and Clostridia were highly abundant, reflecting inflammatory and anaerobic metabolic environments. Orders Enterobacterales and Bacteroidales were dominant, consistent with CD-associated microbial shifts. Family-level analysis showed an imbalance between pro-inflammatory Enterobacteriaceae and protective Lachnospiraceae.

Notably, a significant proportion of sequences remained unclassified at the genus (∼30%) and species (∼50%) levels, highlighting potential novel taxa or database limitations.

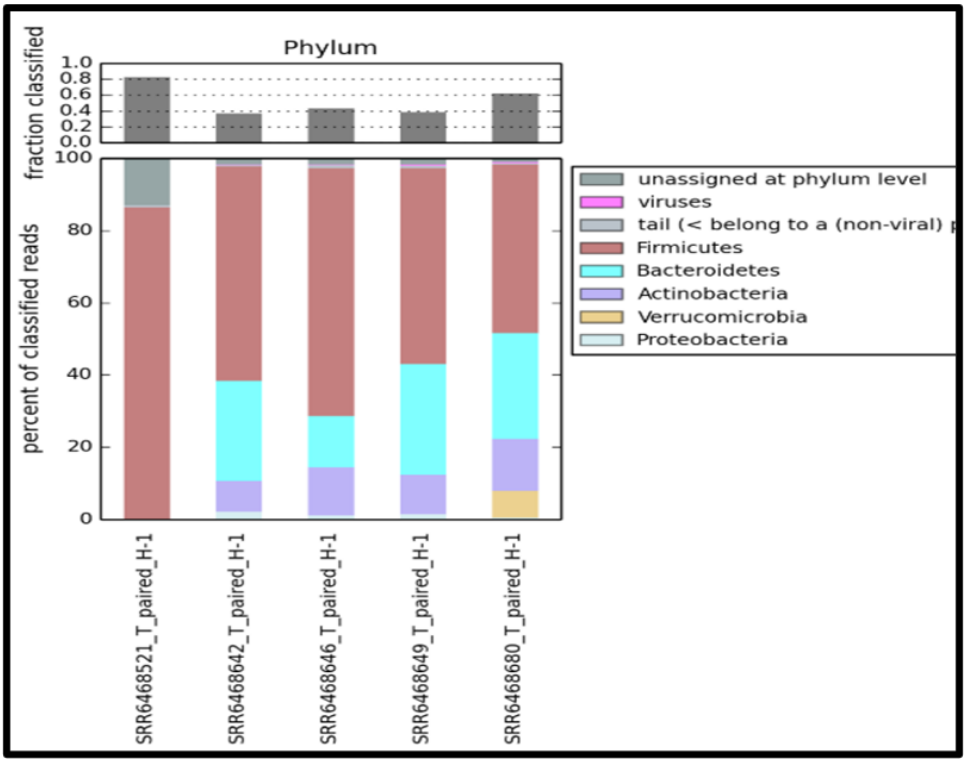

FIGURE 4.17: Classification of Health Control Sample at Phylum level

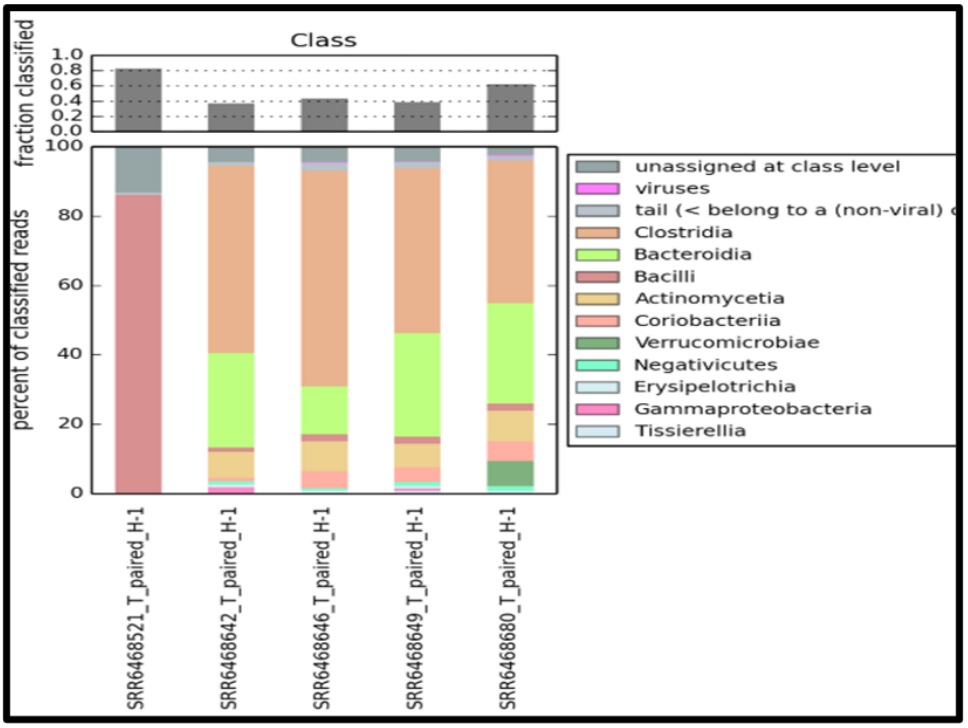

FIGURE 4.18: Classification of Health Control Sample at Class level

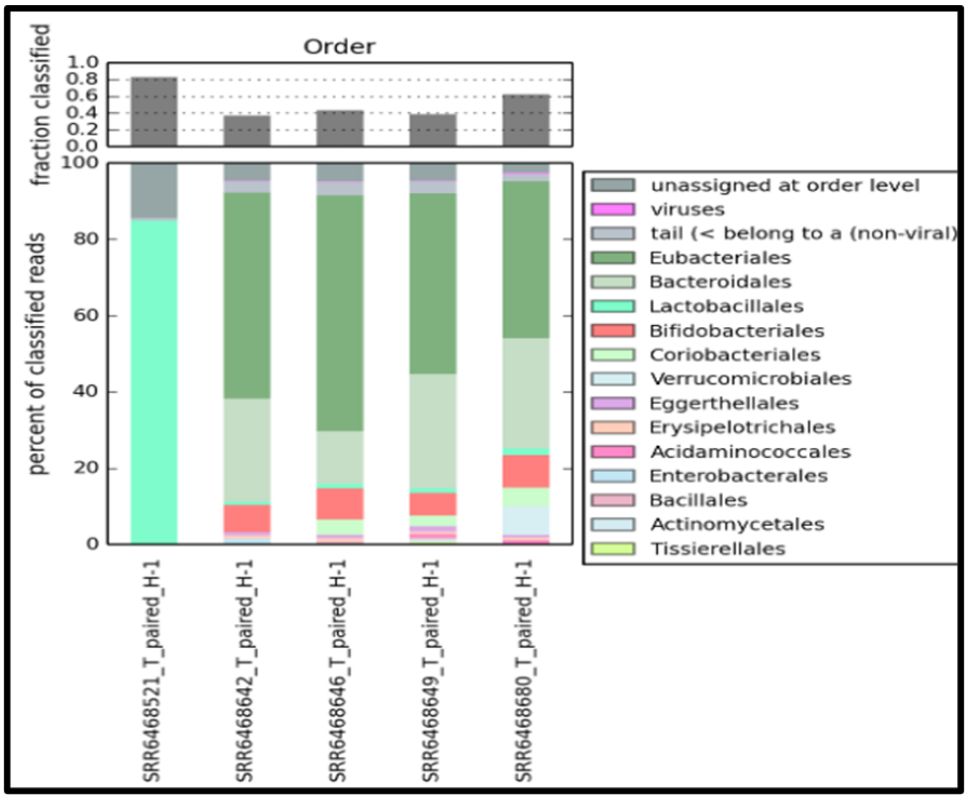

FIGURE 4.19: Classification of Health Control Sample at Order level

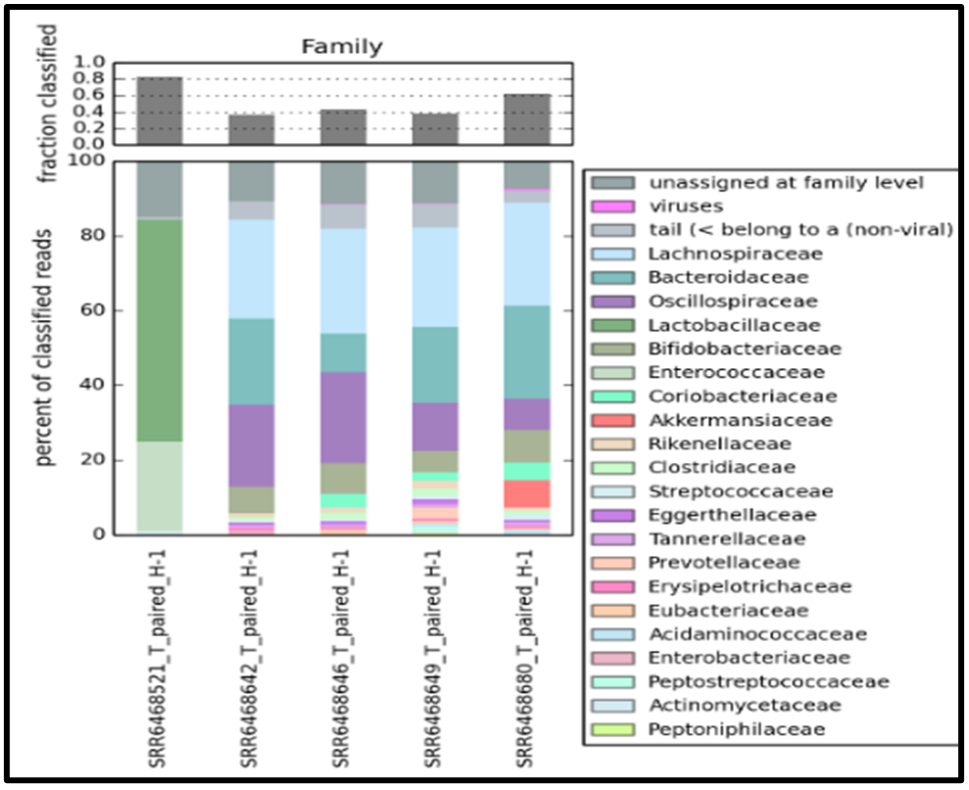

FIGURE 4.20: Classification of Health Control Sample at Family level

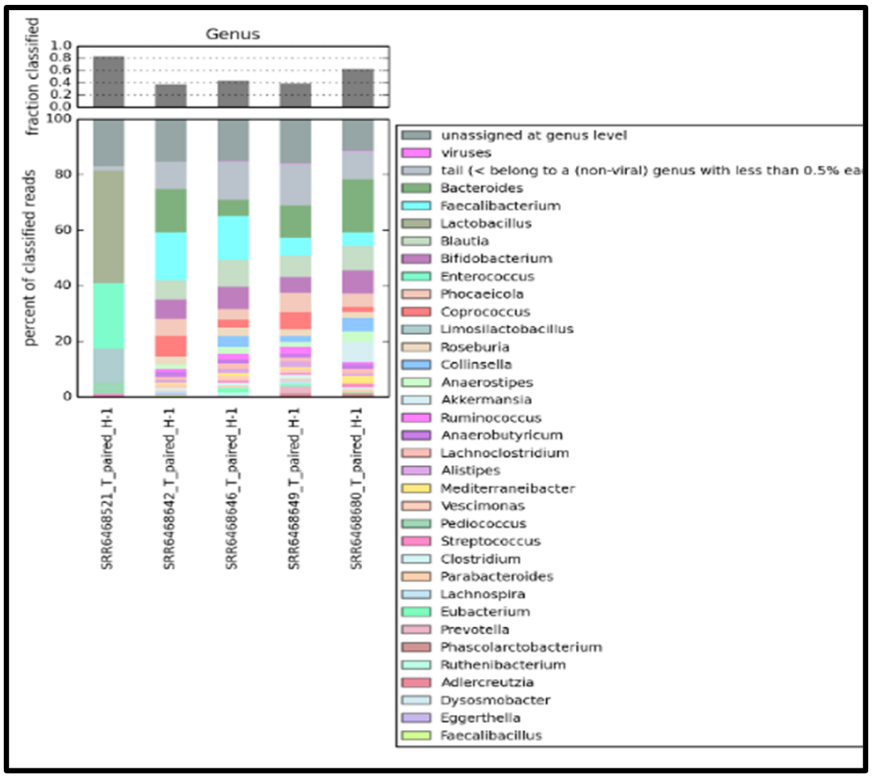

FIGURE 4.21: Classification of Health Control Sample at Genus level

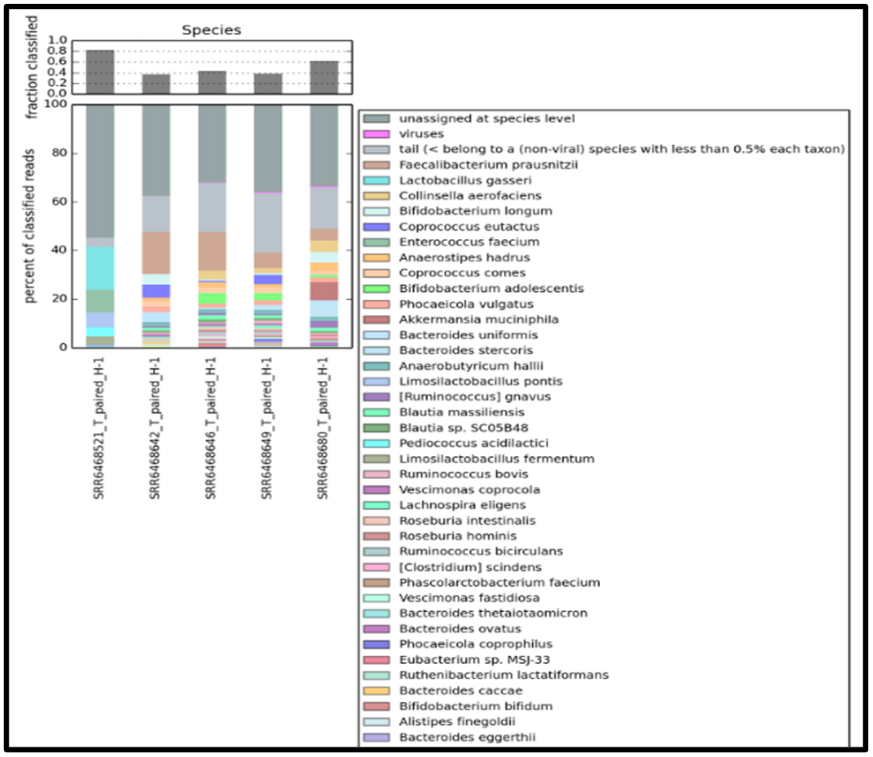

FIGURE 4.22: Classification of Health Control Sample at Species level

The gut microbiome analysis revealed distinct taxonomic compositions across multiple levels. Firmicutes and Bacteroidetes were the most abundant phyla, essential for digestion, immune modulation, and gut barrier integrity. Actinobacteria (including *Bifidobacterium*) and Proteobacteria were present in smaller proportions, while *Verrucomicrobia* (*Akkermansia muciniphila*) contributed to gut health. At the class level, *Clostridia* (Firmicutes) and *Bacteroidia* (Bacteroidetes) dominated, with *Bacilli* (Lactobacillus, Streptococcus) and *Actinomycetia* (Bifidobacterium) also present. Bacteroidales and Eubacteriales were the predominant orders, with beneficial groups such as *Lactobacillales* and *Bifidobacteriales* playing key roles in gut homeostasis.At the family level, *Lachnospiraceae* and *Bacteroidaceae* were prominent, known for producing anti-inflammatory short-chain fatty acids (SCFAs) like butyrate. Other families, including *Bifidobacteriaceae*, *Lactobacillaceae*, and *Akkermansiaceae*, contributed to gut stability.Key genera included Bacteroides, Blautia, Lactobacillus, Bifidobacterium, Roseburia, and Akkermansia, playing roles in fiber fermentation, probiotic effects, and gut barrier function. Beneficial species such as *Faecalibacterium prausnitzii* (butyrate producer), *Akkermansia muciniphila* (gut barrier enhancer), *Bifidobacterium longum*, and *Roseburia intestinalis* were identified, alongside commensals like *Bacteroides thetaiotaomicron* and *Phocaeicola vulgatus*, supporting polysaccharide metabolism.

4.4 Multivariable Associations Analysis

4.4.1 Associations Between Microbial Taxa and Diagnosis

Microbial taxa analysis revealed strong positive associations with Crohn's Disease (CD), particularly for *Alistipes indistinctus* (coef = 1.00, qval = 4.58e-60) and *Prevotella copri* (coef = 1.00, qval = 4.62e-60), both of which were significantly enriched in CD patients compared to healthy controls. Additionally, *Fusobacterium nucleatum* (coef = 1.00, qval = 2.26e-59) demonstrated a robust positive correlation with CD, suggesting its potential role in disease pathogenesis.Moderate positive associations were observed for *Ruminococcus gnavus* (coef = 4.13, qval =

0.036) and *Bacteroides fragilis* (coef = 0.94, qval = 0.057), indicating their relative enrichment in CD patients. While their associations were not as strong as the taxa above, their presence suggests a shift in microbial composition associated with CD. These findings highlight specific bacterial species that may serve as potential biomarkers for CD, reflecting disease-specific dysbiosis and alterations in gut microbiome composition.

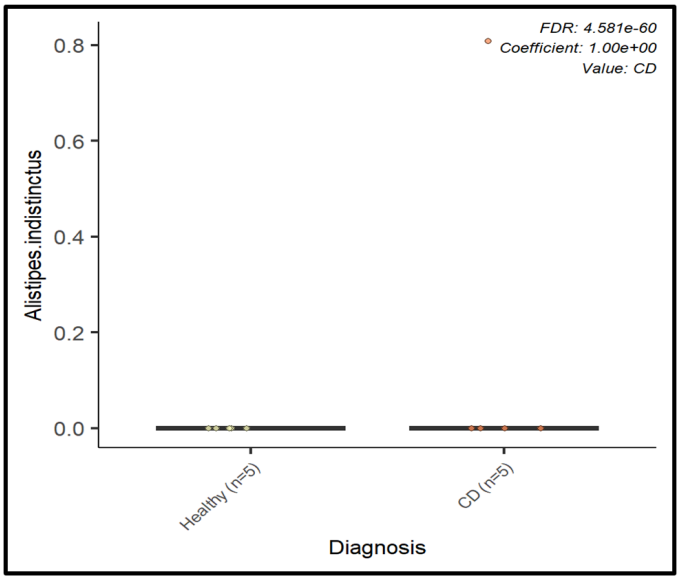

FIGURE 4.23: Strong Positive Associations between diagnosis and *Alistipes indistinctus*

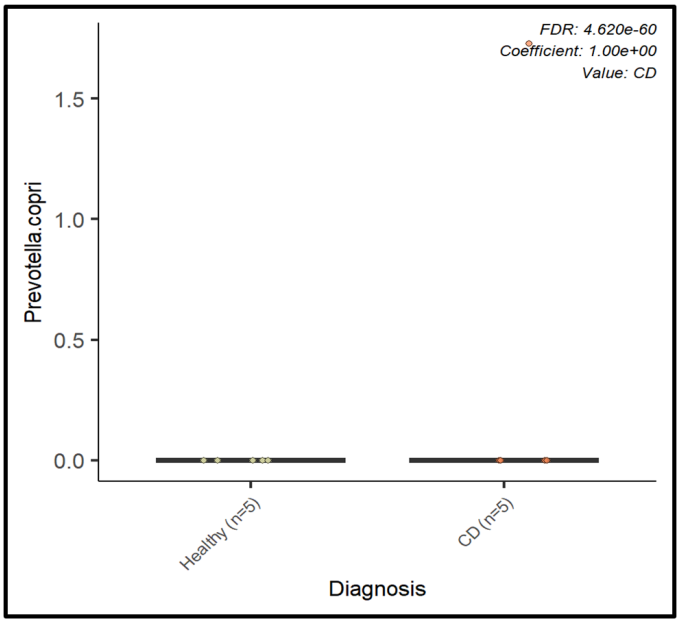

FIGURE 4.24: Strong Positive Associations between diagnosis and *Prevotella copri*

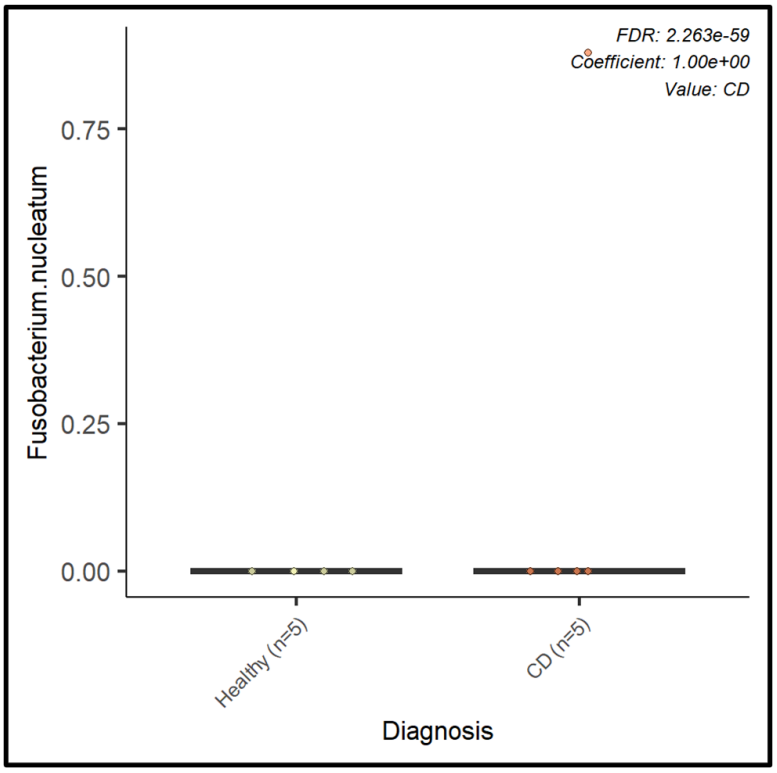

FIGURE 4.25: Strong Positive Associations between diagnosis and *Fusobacterium nucleatum*

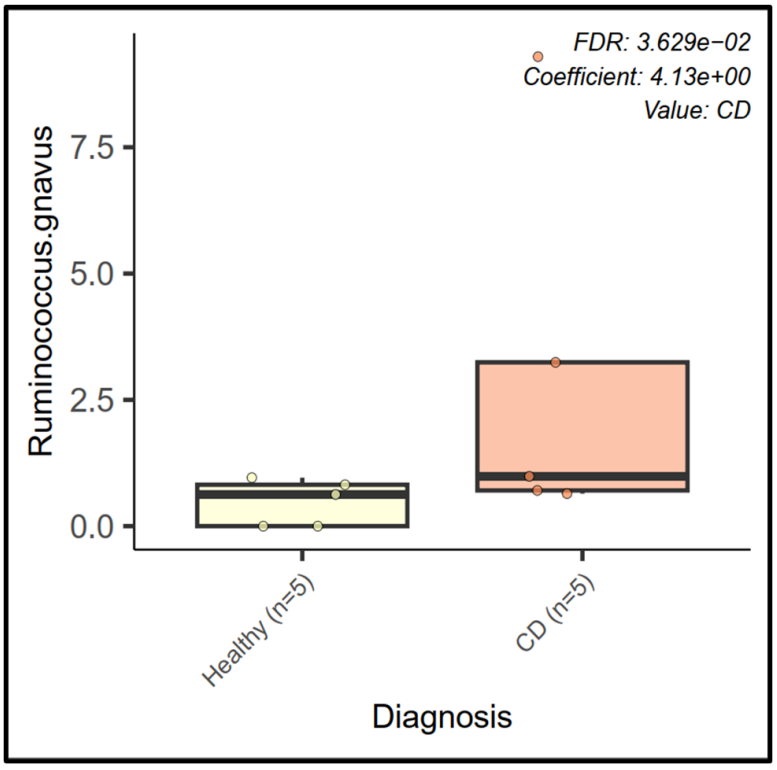

FIGURE 4.26: Moderate Positive Associations between diagnosis and *Ruminococcus gnavus*

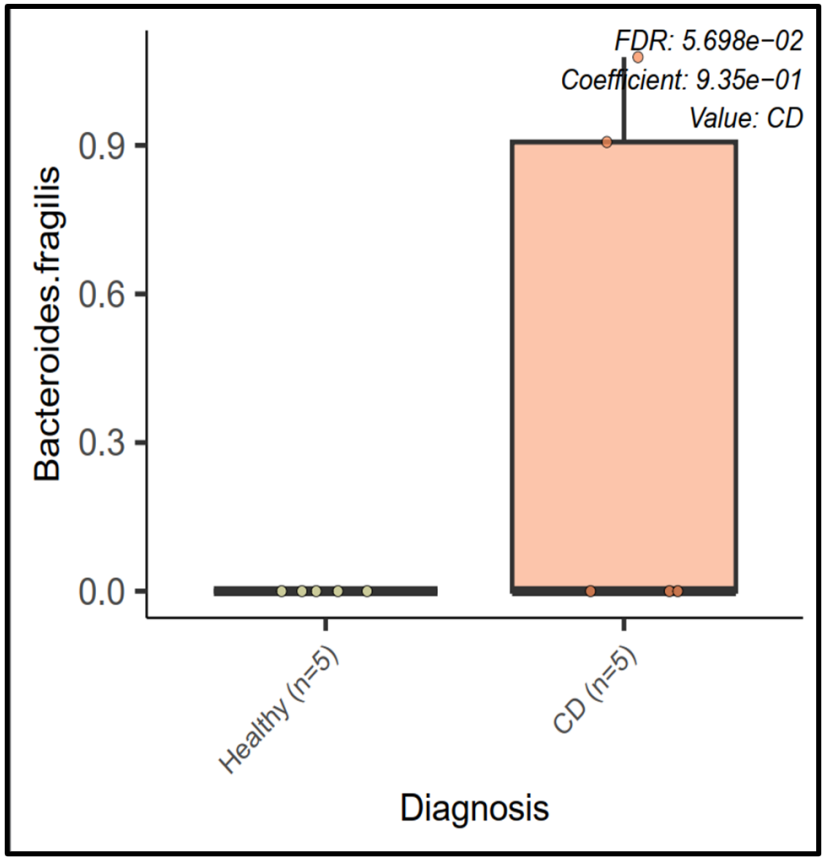

FIGURE 4.27: Moderate Positive Associations between diagnosis and *Bacteroides fragilis*

4.4.2 Associations Between Microbial Taxa and Age

The analysis revealed notable associations between microbial taxa and age. *Parabacteroides merdae* (coef = -0.46, qval = 0.063) and *Blautia hansenii* (coef = -0.26, qval = 0.063) exhibited a negative correlation with age, indicating a decline in their abundance as individuals get older. In contrast, *Phocaeicola dorei* (coef = 0.26, qval = 0.063) showed a positive association with age, suggesting an increase in its abundance over time.

These findings suggest that age-related shifts in the gut microbiome may play a role in microbial dysbiosis, potentially influencing gut health in older individuals.

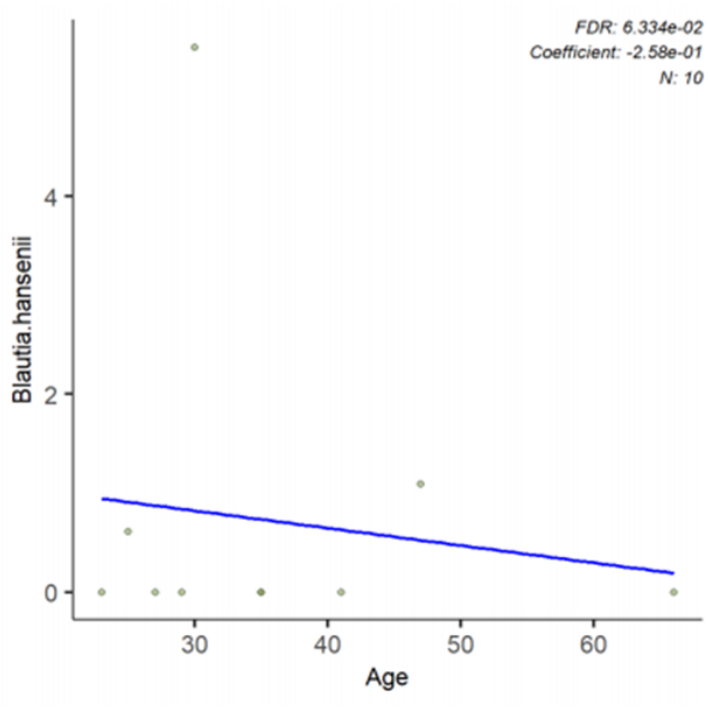

FIGURE 4.28: Negative association between age and *Blautia hansenii*

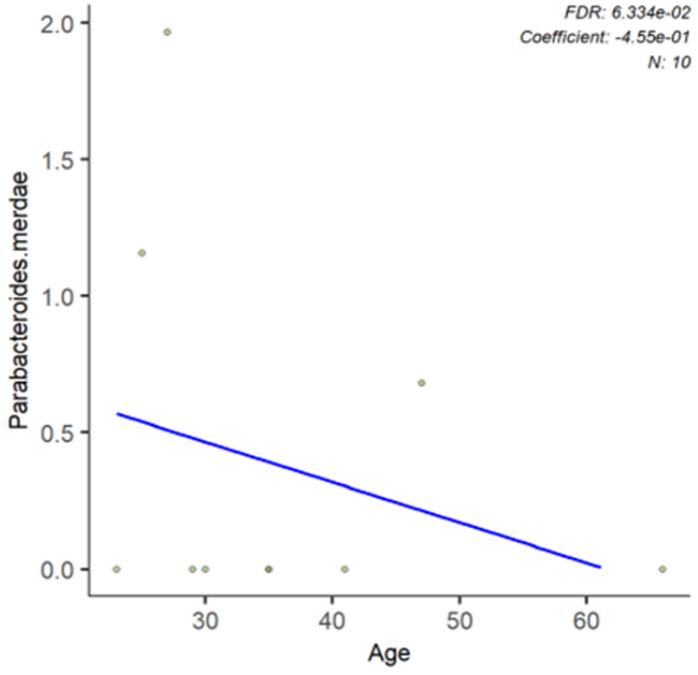

FIGURE 4.29: Negative association between age and *Parabacteroides merdae*

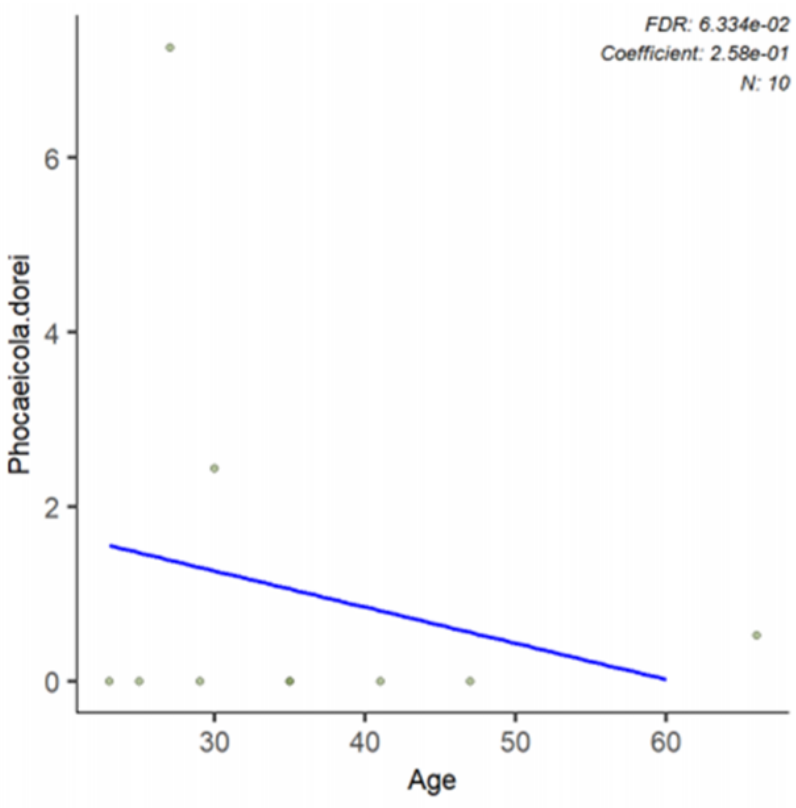

FIGURE 4.30: Positive Association was found between Age and *Phocaeicola dorei*

4.4.3 Associations Between Microbial Taxa and Mesalamine Usage

The analysis identified significant microbial associations with mesalamine treatment. *Alistipes onderdonkii* (coef = 0.59, qval = 4.58e-60) and *Phocaeicola dorei* (coef = 1.65, qval = 0.00016) were positively associated with mesalamine use, indicating their enrichment in patients undergoing this treatment. Conversely, *Collinsella stercoris* (coef = -0.32, qval = 2.26e-59) and *Bifidobacterium pseudocatenulatum* (coef = -0.32, qval = 1.13e-58) exhibited a negative association, suggesting a reduction in their abundance with mesalamine therapy.

These findings highlight the potential impact of mesalamine on gut microbiome composition, which may influence its therapeutic effects in inflammatory bowel disease management.

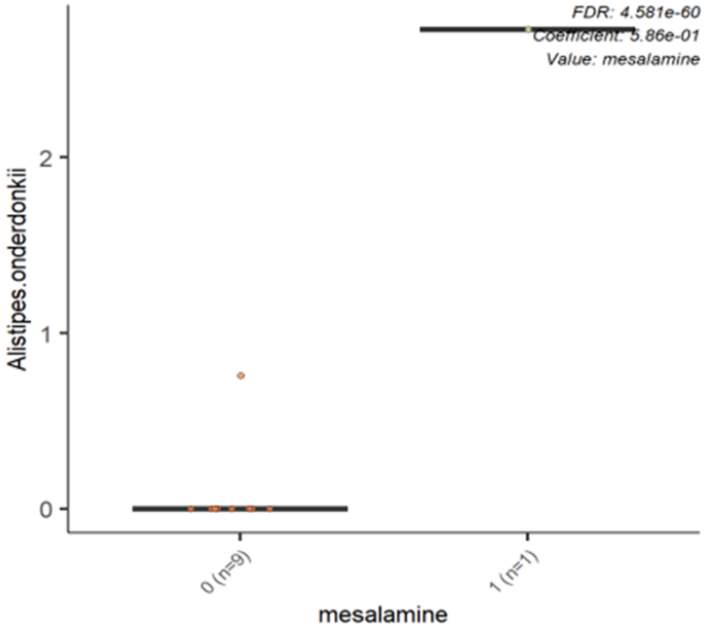

FIGURE 4.31: Positive Associations between Mesalamine usage and *Alistipes onderdonkii*

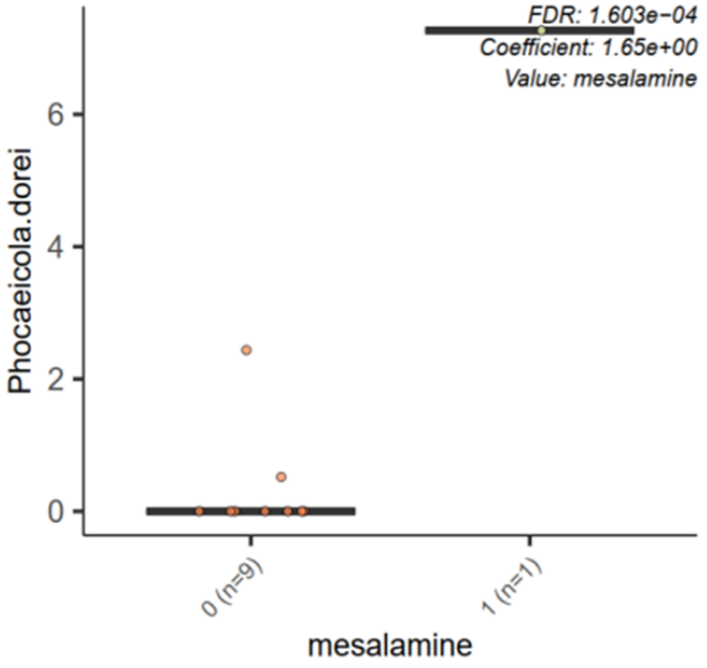

FIGURE 4.32: Positive Associations between Mesalamine usage and *Phocaeicola dorei*

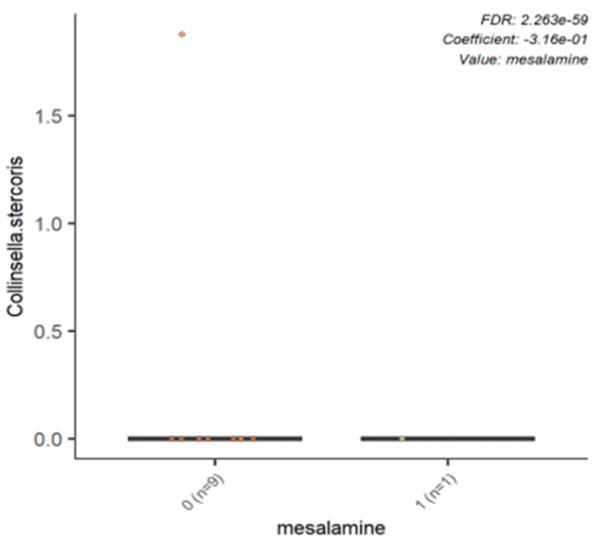

FIGURE 4.33: Negative Associations between Mesalamine usage and *Collinsel stercoris*

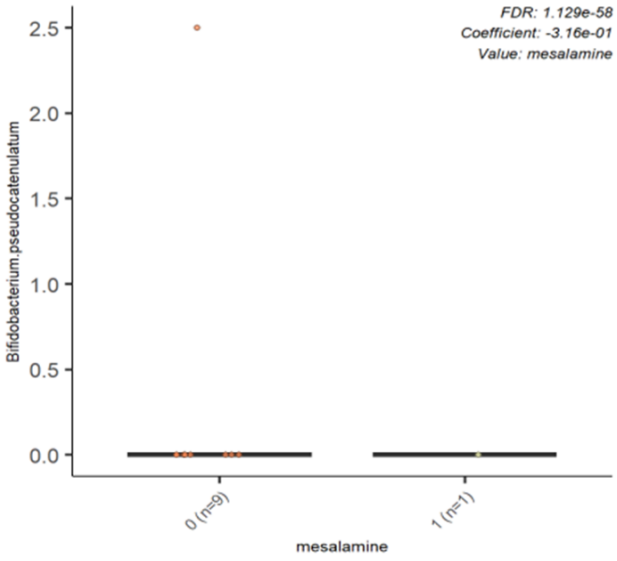

FIGURE 4.34: Negative Associations between Mesalamine usage and *Bifidobacterium pseudocatenulatum*

4.4.4 Steroid Use and Microbial Shifts

The analysis revealed significant associations between microbial taxa and steroid treatment. Negative associations were observed for *Alistipes indistinctus* (coef = -0.42, qval = 4.58e-60) and *Clostridium innocuum* (coef = -0.42, qval = 9.22e-60), indicating that these taxa were less abundant in individuals receiving steroid

treatment. Conversely, positive associations were found for *Collinsella stercoris* (coef = 0.42, qval = 2.19e-59) and *Bifidobacterium pseudocatenulatum* (coef = 0.42, qval = 1.06e-58), suggesting that these species were enriched in steroid-treated individuals. These findings imply that steroid therapy may suppress inflammation-associated bacteria while promoting taxa that could play a role in gut homeostasis, potentially influencing therapeutic outcomes in Crohn's Disease management.

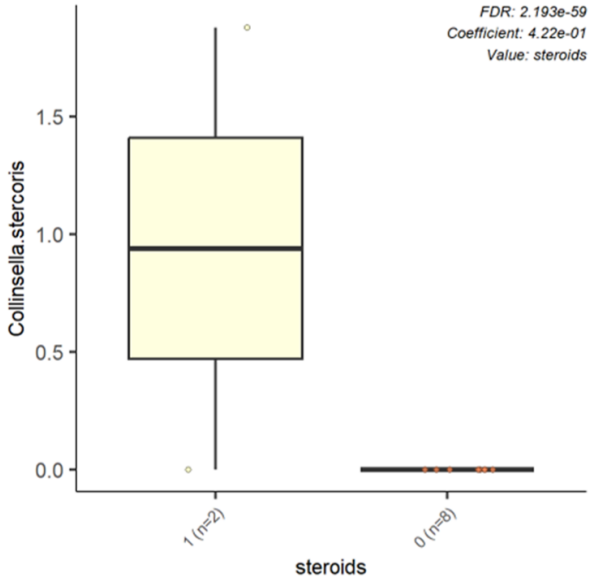

FIGURE 4.35: Positive Associations of Steroid usage and *Collinsella stercoris*

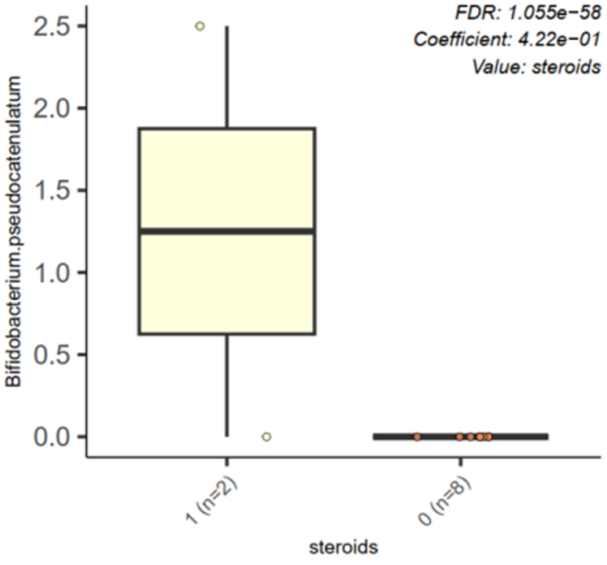

FIGURE 4.36: Positive Associations of Steroid usage and *Bifidobacterium pseudocatenulatum*

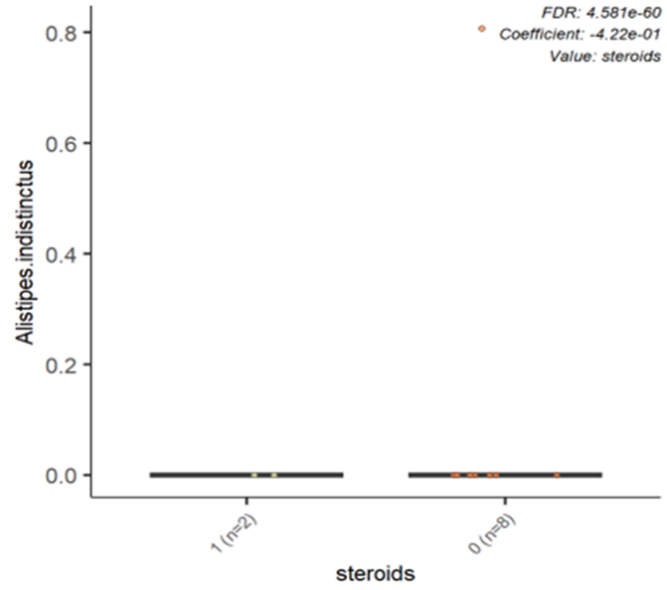

FIGURE 4.37: Negative Associations of Steroid usage and *Alistipes indistinctus*

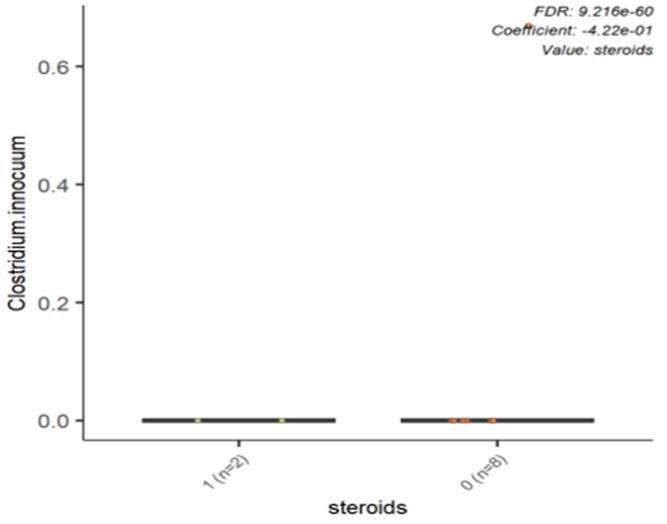

FIGURE 4.38: Negative Associations of Steroid usage and *Clostridium innocuum*

4.4.5 Associations Between Microbial Taxa and Immunosuppressant Treatment

The use of immunosuppressants was associated with significant reductions in certain microbial taxa. Strong negative associations were observed for *Alistipes indistinctus* (coef = -0.42, qval = 4.58e-60), *Prevotella copri* (coef = -0.42, qval =

6.81e-60), and *Blautia hansenii* (coef = -1.41, qval = 0.00047), indicating that these species were significantly depleted in individuals undergoing immunosuppressant therapy.

Additionally, moderate negative associations were noted for *Clostridium hylemonae* (coef = -0.53, qval = 8.08e-59) and *Ruminococcus gnavus* (coef = -1.69, qval = 0.065), suggesting a trend toward reduced abundance in treated individuals. These findings suggest that immunosuppressants may influence gut microbial composition by suppressing taxa linked to inflammation, potentially affecting microbial homeostasis and overall gut health in Crohn's Disease patients.

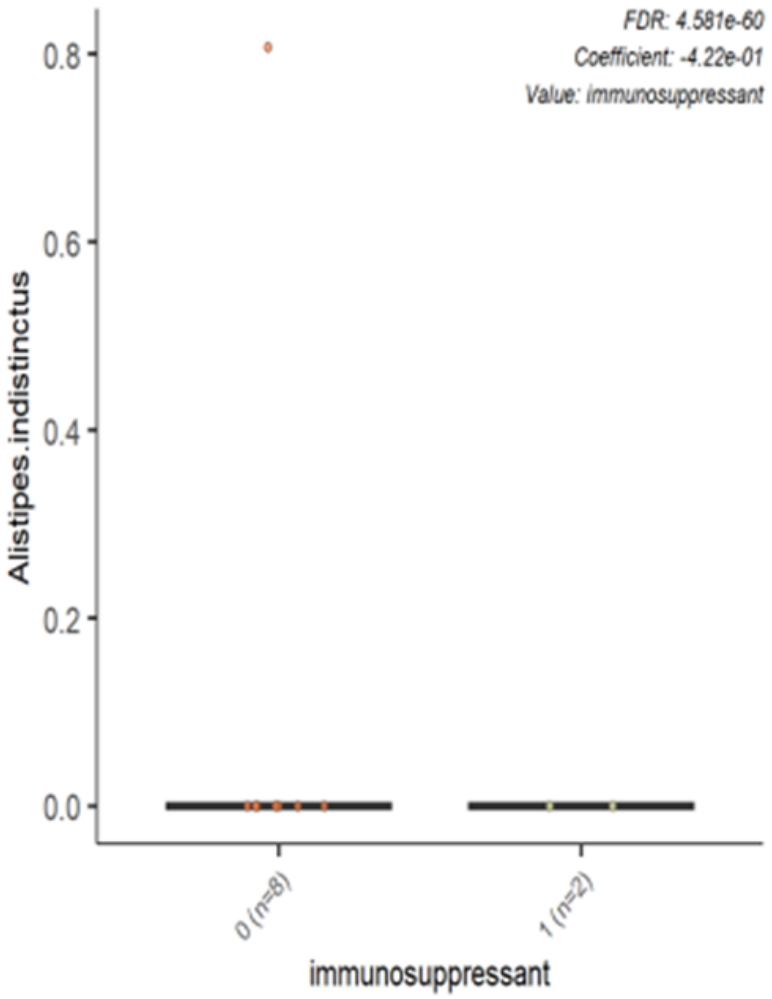

FIGURE 4.39: Negative Associations of Immunosuppressants with *Alistipes indistinctus*

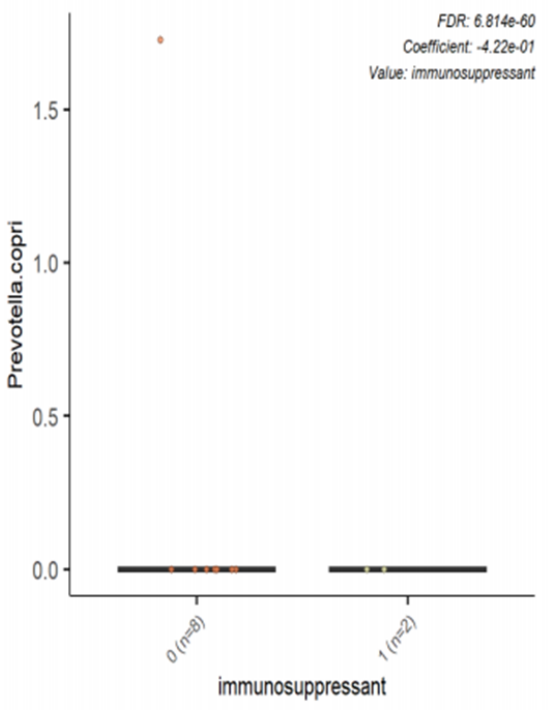

FIGURE 4.40: Negative Associations of Immunosuppressants with *Prevotella copri*

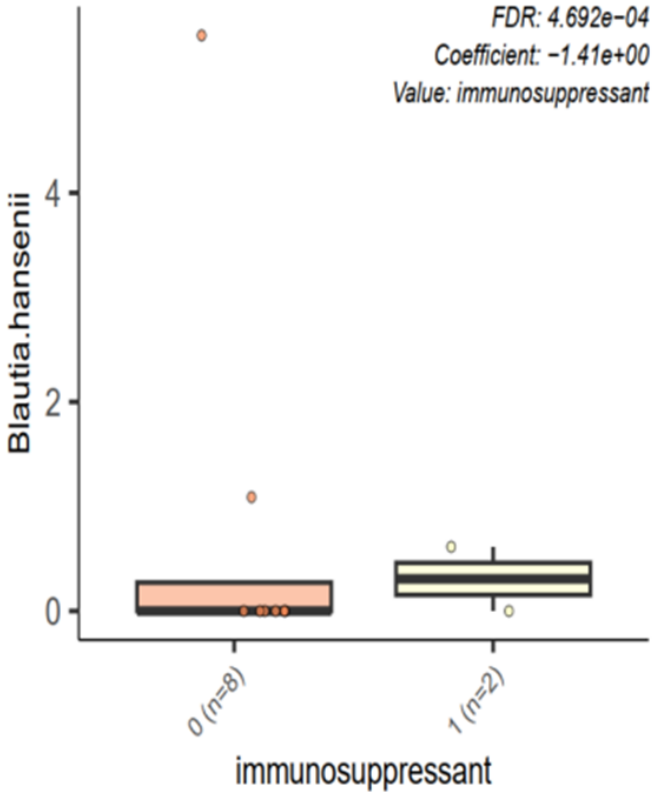

FIGURE 4.41: Negative Associations of Immunosuppressants with *Blautia hansenii*

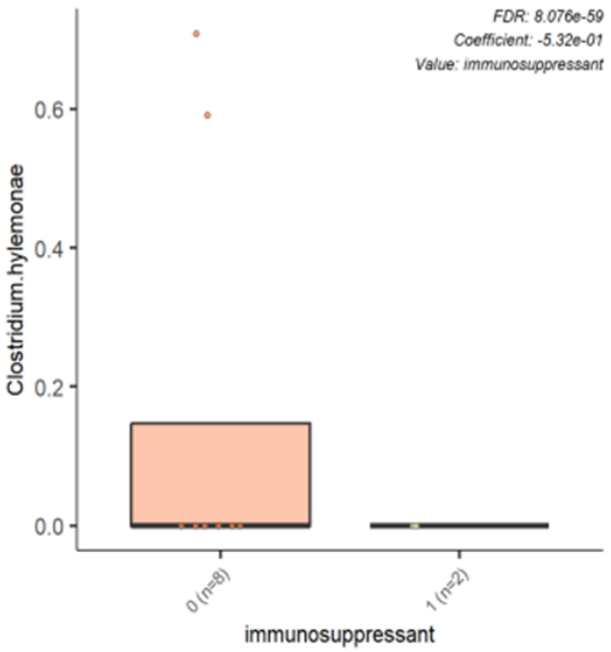

FIGURE 4.42: Moderate Negative Associations of Immunosuppressants with *Clostridium hylemonae*

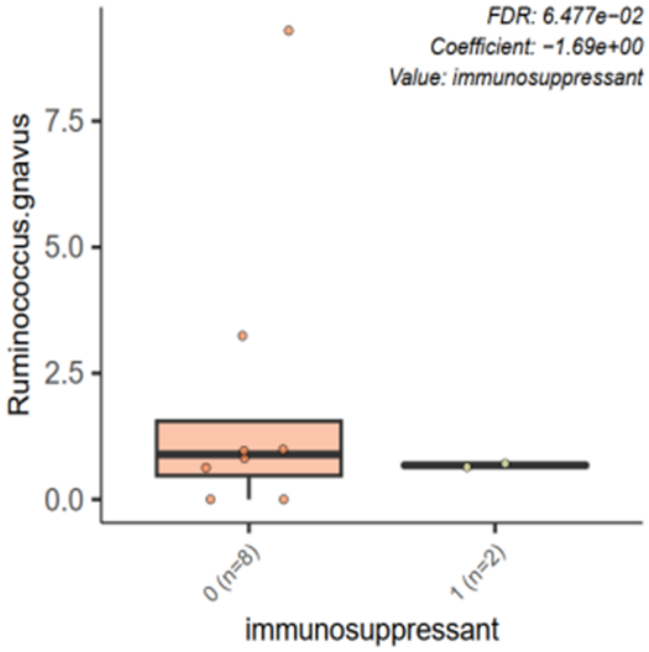

FIGURE 4.43: Moderate Negative Associations of Immunosuppressants with *Ruminococcus gnavus*

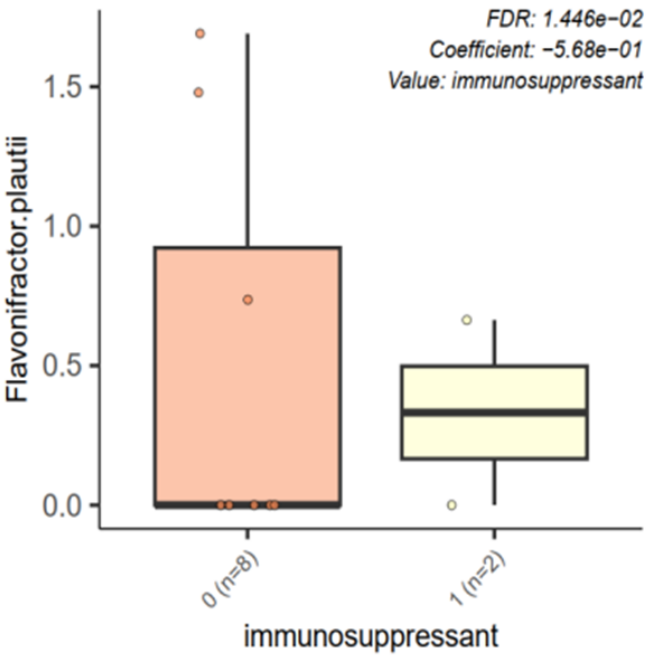

FIGURE 4.44: Moderate Negative Associations of Immunosuppressants with *Flavonifractor plautii*

4.5 Visualization of Microbial Taxa Associations

The heatmap visualization highlights the top 50 microbial taxa significantly associated with key clinical variables, including Crohn's Disease (CD), immunosuppressants, steroids, mesalamine treatment, and age. Strong positive associations (red) were observed for *Alistipes indistinctus* and *Prevotella copri*, both enriched in CD patients, suggesting their potential as disease biomarkers. Additionally, *Phocaeicola dorei* was positively associated with mesalamine treatment, indicating a possible role in therapeutic response. In contrast, strong negative associations (blue) were noted for *Blautia pseudococcoides* and *Clostridium innocuum*, which were significantly depleted with immunosuppressant use, while *Collinsella stercoris* showed a marked reduction in response to steroid treatment. The heatmap effectively clusters taxa based on their responses to clinical factors, providing a clear visual representation of microbial shifts that may aid in understanding disease mechanisms and treatment effects.

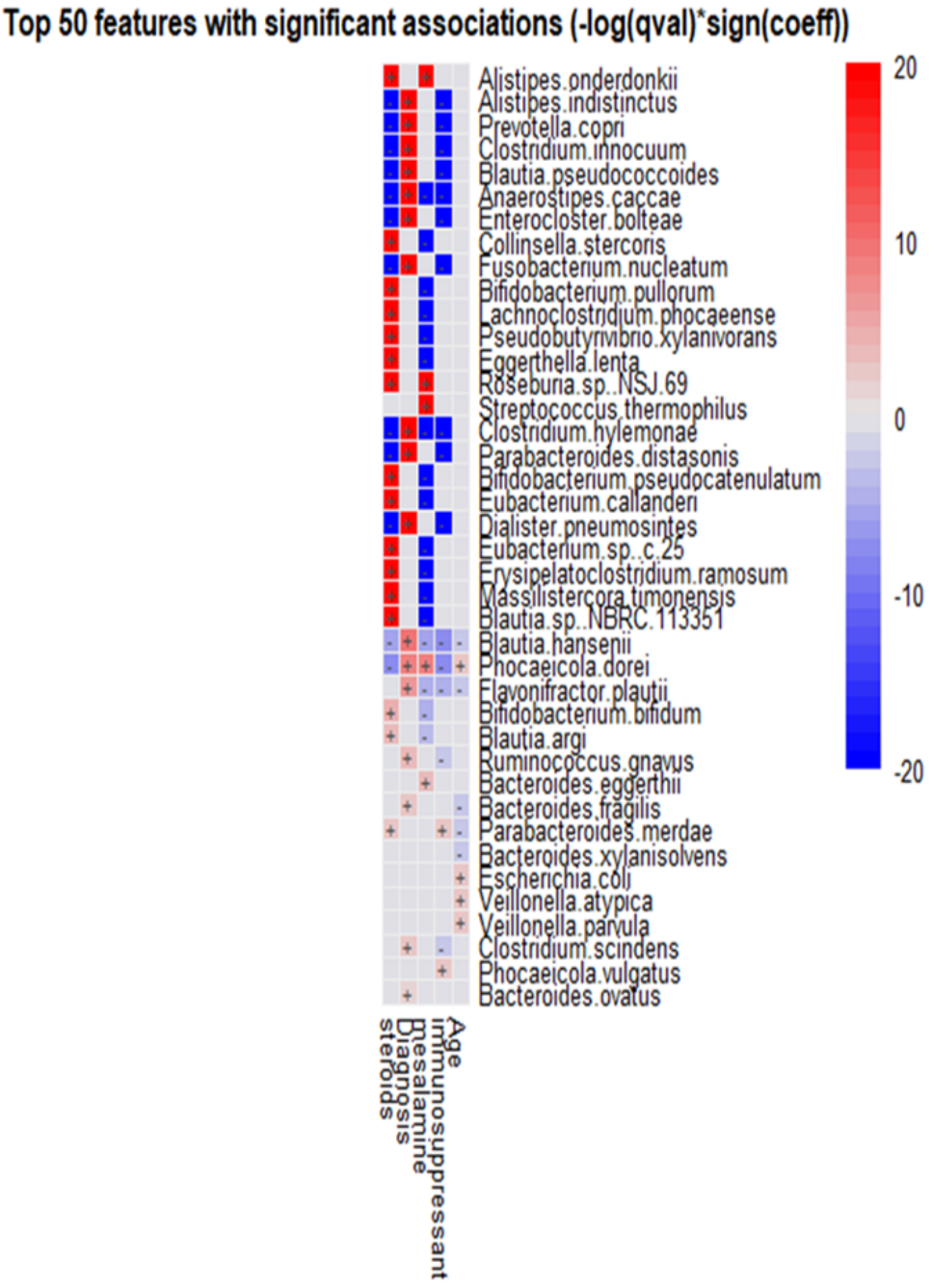

FIGURE 4.45: Heatmap visualization highlighting tthe top 50 microbial taxa significantly associated with Crohn's Disease

4.5.1 Visualizations

Krona Chart Krona visualizations were generated for each individual sample in both the Crohn's Disease (CD) and Healthy groups, allowing for an interactive exploration of the microbial community composition. These circular, multi-layered graphs provide a hierarchical representation of taxonomic classification, enabling

detailed insights into the relative abundance of different microbial taxa at various taxonomic levels (phylum, genus, species).For CD samples, the Krona plots reveal a distinct microbial composition, often characterized by an enrichment of taxa previously associated with gut dysbiosis and inflammation. In contrast, Healthy samples display a more balanced microbial profile, with a higher prevalence of beneficial commensals. The ability to drill down into each taxonomic level within Krona allows for a comprehensive comparison of microbial diversity across individual samples, highlighting the structural differences in gut microbiota between diseased and healthy states. These visualizations serve as an essential tool for identifying microbial shifts at an individual level and their potential role in disease progression.

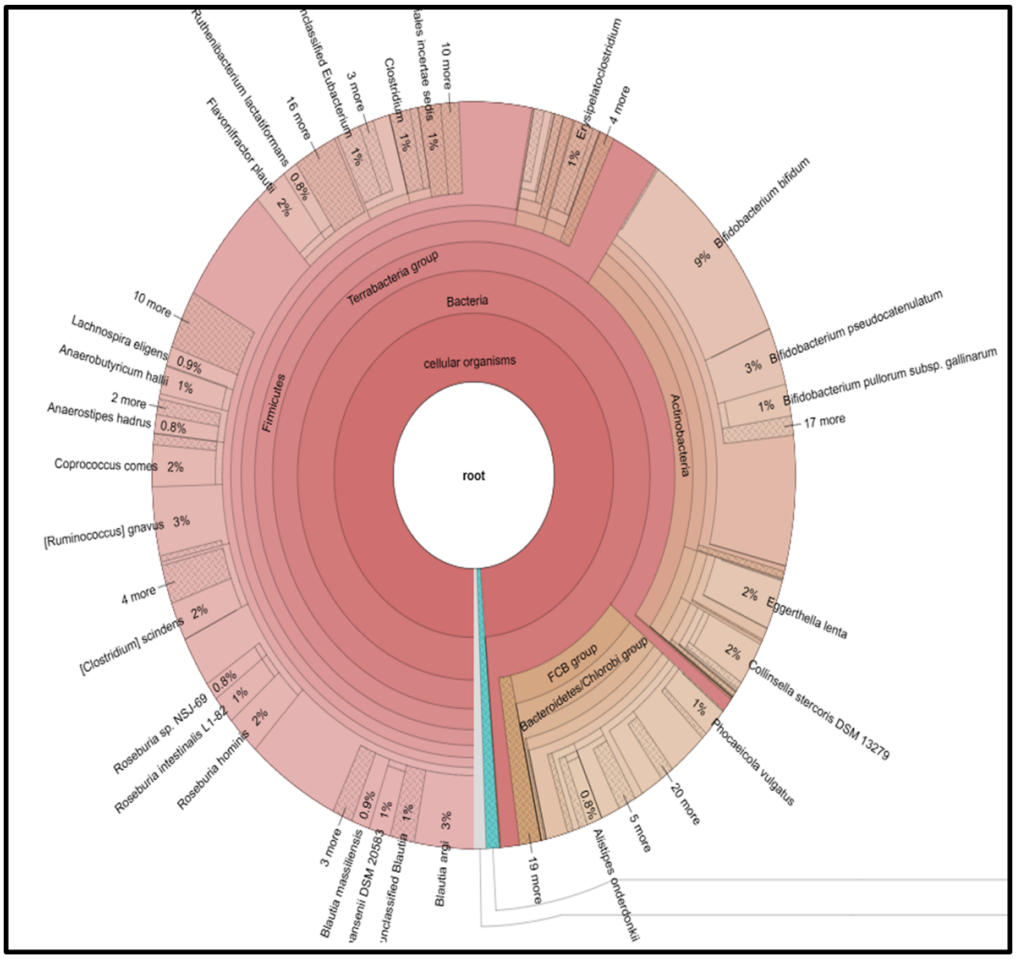

FIGURE 4.46: Krona Chart Visualization of bacterial species from Crohn's disease patient samples with ID SRR6468520

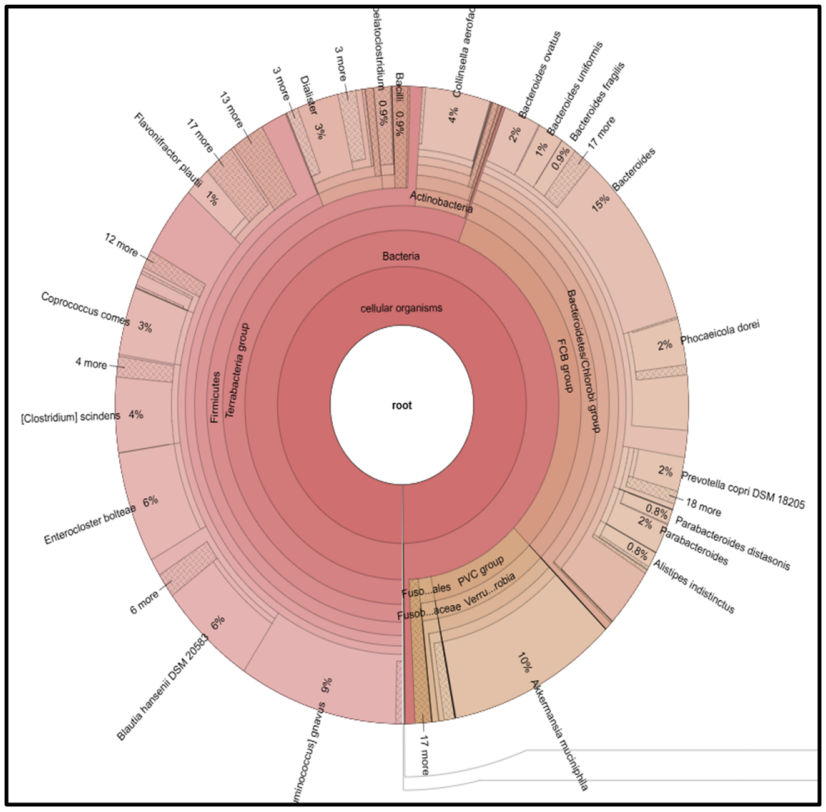

FIGURE 4.47: Krona Chart Visualization of bacterial species from Crohn's Disease patient samples with ID SRR6468527

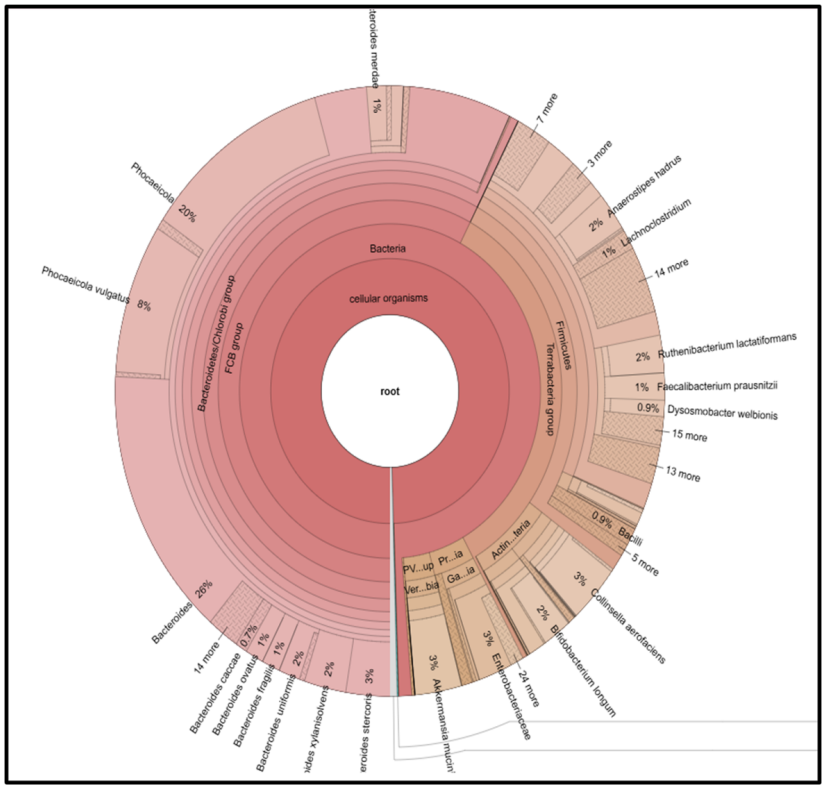

FIGURE 4.48: Krona Chart Visualization of bacterial species from Crohn's Disease patient samples with ID SRR6468559

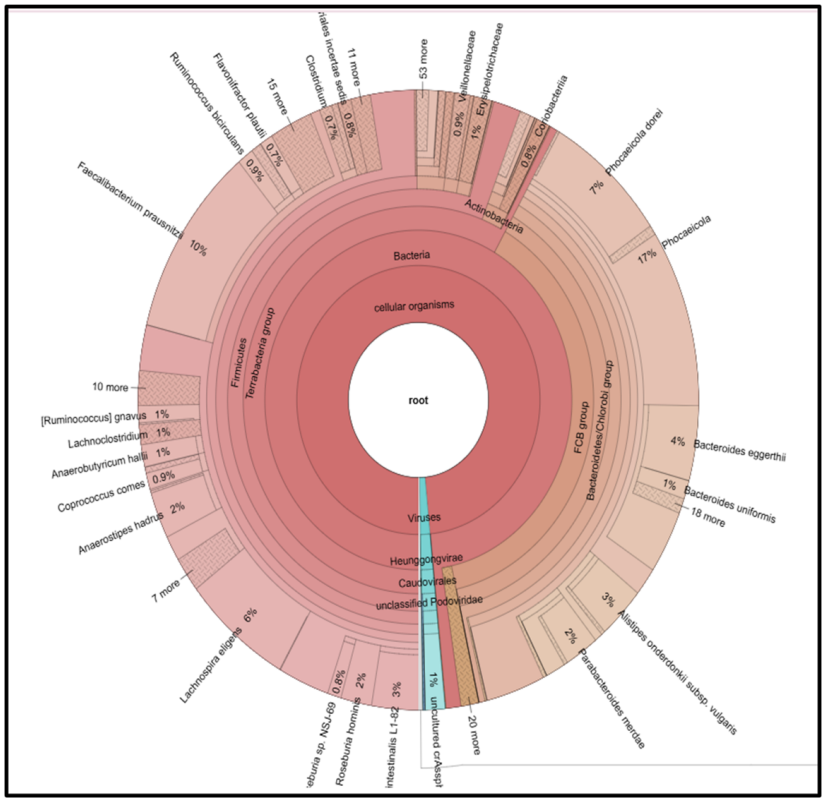

FIGURE 4.49: Krona Chart Visualization of bacterial species from Crohn's Disease patient samples with ID SRR6468560

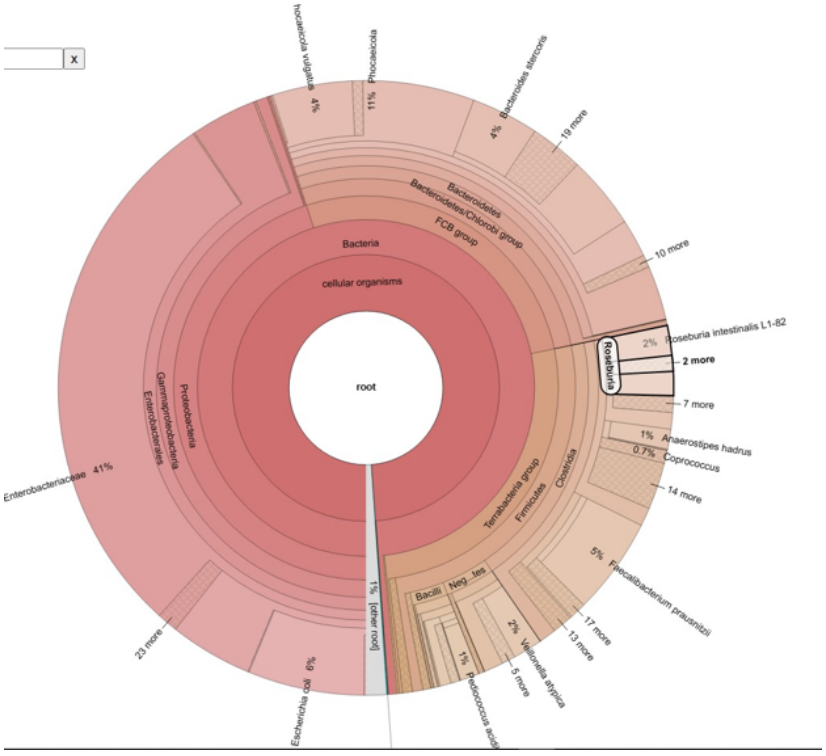

FIGURE 4.50: Krona Chart Visualization of bacterial species from Crohn's Disease patient samples with ID SRR6468561

FIGURE 4.51: Krona Chart Visualization of bacterial species from Healthy Control samples with ID SRR6468521

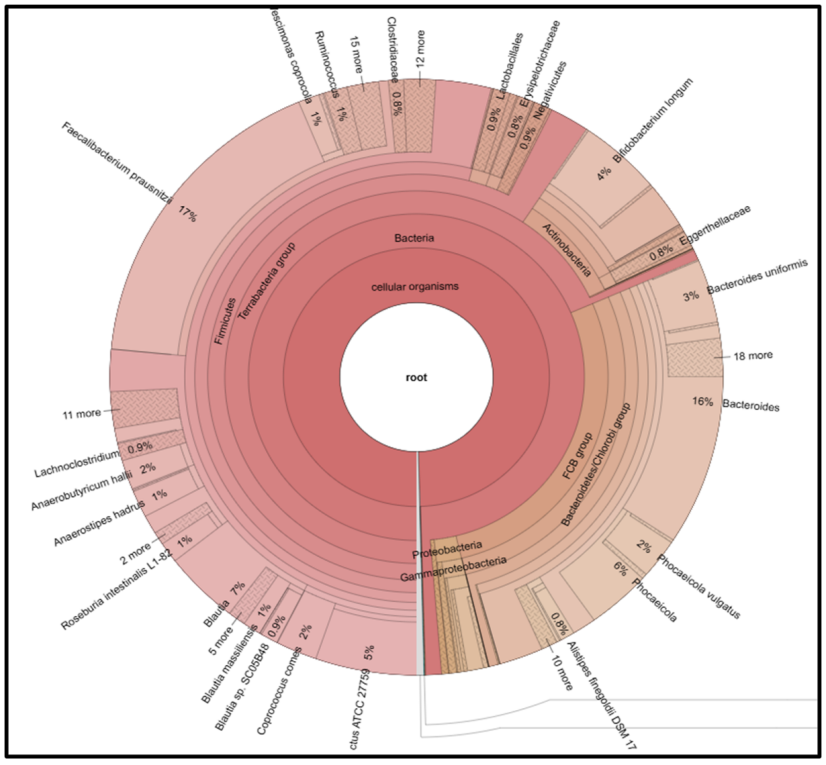

FIGURE 4.52: Krona Chart Visualization of bacterial species from Healthy Control samples with ID SRR6468642

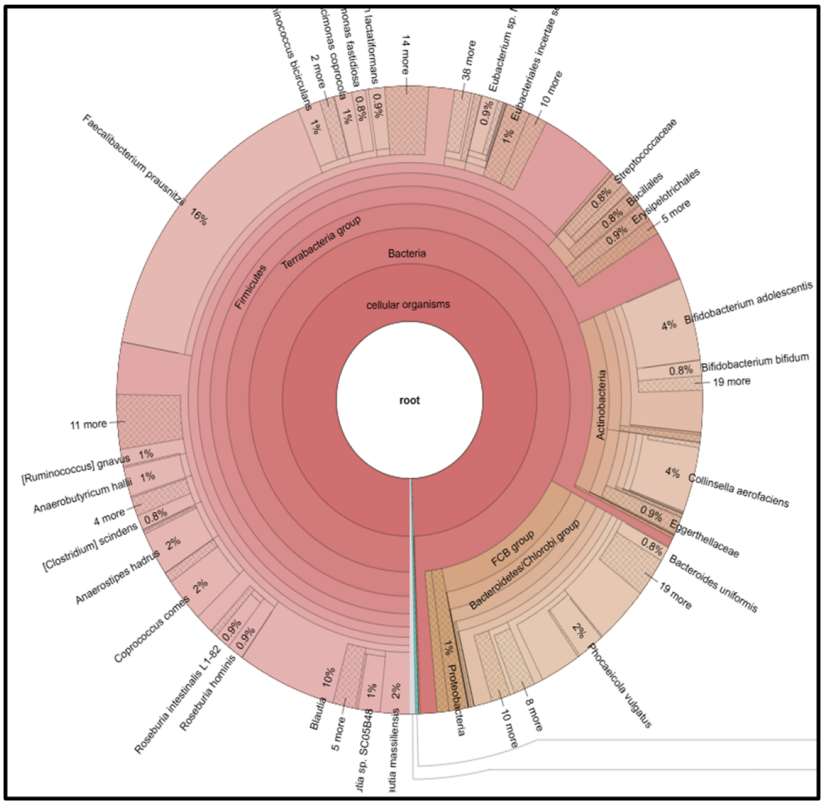

FIGURE 4.53: Krona Chart Visualization of bacterial species from Healthy Control samples with ID SRR6468646

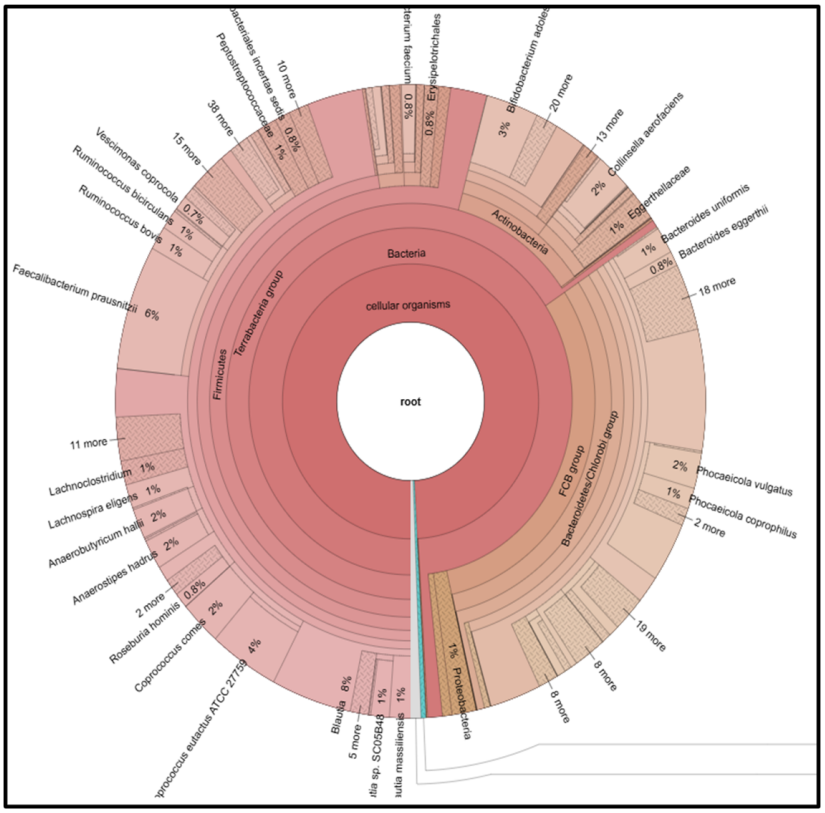

FIGURE 4.54: Krona Chart Visualization of bacterial species from Healthy Control samples with ID SRR6468649

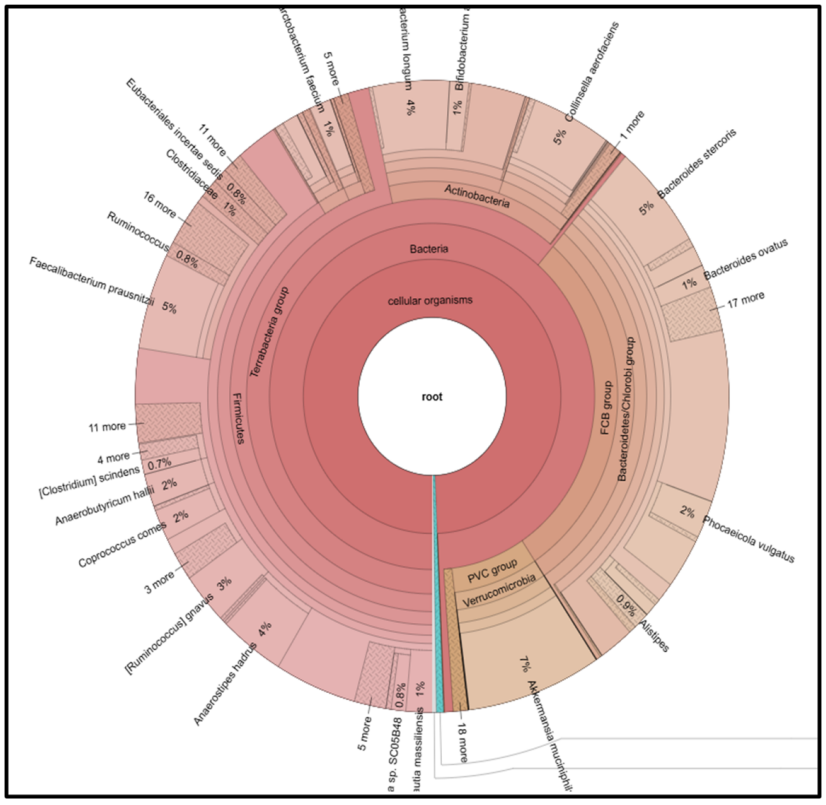

FIGURE 4.55: Krona Chart Visualization of bacterial species from Healthy Control samples with ID SRR6468680

4.6 Comparison of Taxon Abundance in CD VS Healthy

The heatmap highlights microbial shifts in Crohn's Disease (CD), with *Ruminococcus gnavus*, *Enterococcus faecium*, and *Bacteroides stercoris* enriched in CD, indicating inflammation and dysbiosis. In contrast, beneficial taxa like *Faecalibacterium prausnitzii*, *Roseburia intestinalis*, and *Akkermansia muciniphila* are significantly depleted, reflecting impaired gut barrier integrity and reduced SCFA production. Some taxa, such as *Bifidobacterium adolescentis* and *Lactobacillus gasseri*, show variable trends. The color gradient (positive values indicating higher abundance in healthy samples and negative values in CD) underscores a loss of protective bacteria and an expansion of pro-inflammatory species, aligning with known CD-associated dysbiosis and potential targets for therapeutic intervention.

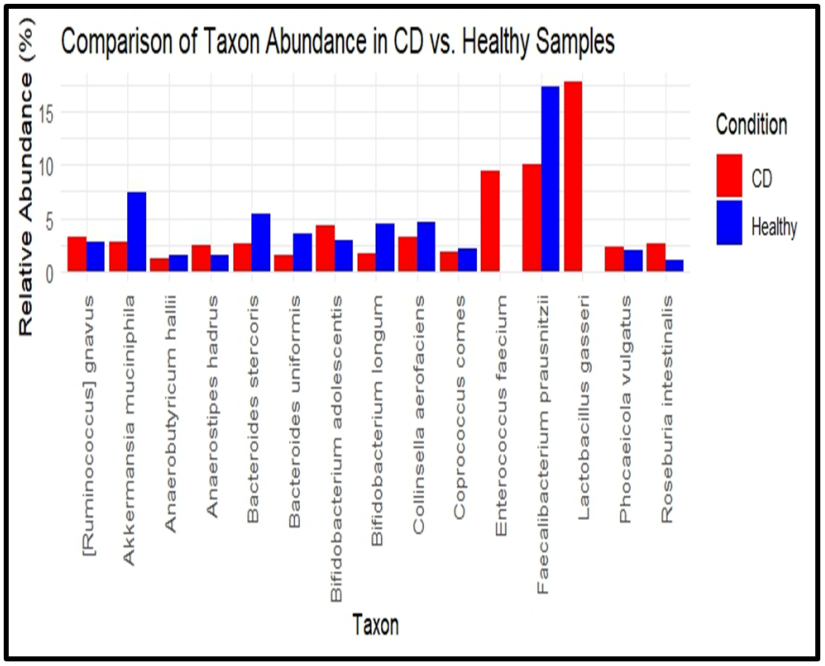

FIGURE 4.56: Comparison of Taxon Abundance in Crohn's Disease Patients (Red) Compared to Healthy Control Samples (Blue)

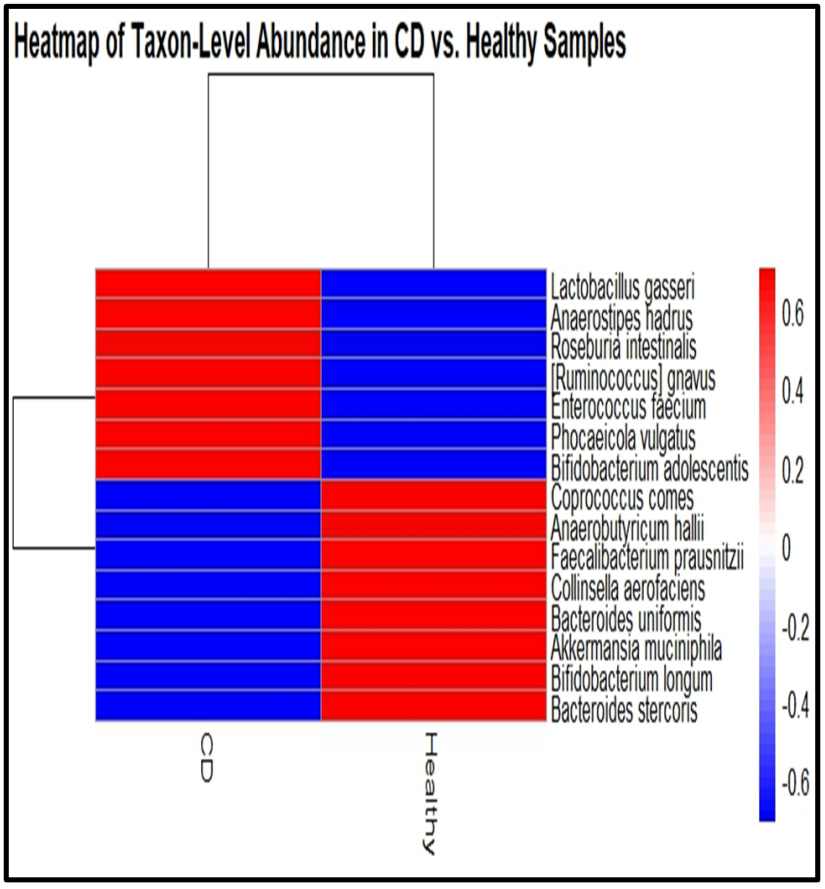

FIGURE 4.57: Heatmap Illustrating Taxon-Level Abundance in Samples from Individuals with Crohn's Disease Compared to Healthy individuals

4.7 Discussion

The gut microbiome in Crohn's Disease (CD) exhibits significant dysbiosis, with an enrichment of pro-inflammatory taxa such as *Alistipes indistinctus*, *Prevotella copri*, and *Ruminococcus gnavus* and a depletion of beneficial taxa like *Faecalibacterium prausnitzii* and *Roseburia intestinalis*. These alterations support the "inflammogenic microbiome" hypothesis, where pathogenic microbes contribute to mucosal inflammation, while anti-inflammatory butyrate producers are suppressed. Therapeutic interventions further modulate microbial composition, as immunosuppressants reduce *Blautia pseudococcoides* and *Clostridium innocuum*, mesalamine promotes *Phocaeicola dorei* and *Alistipes onderdonkii*, and steroids suppress Alistipes while increasing *Collinsella stercoris*, a taxon linked to bile acid metabolism. Age-related changes, including declines in *Parabacteroides merdae* and *Blautia hansenii*, suggest that microbiome shifts could influence CD severity in older individuals. Despite robust statistical adjustments using FDR correction ($q < 0.1$), the study's small cohort size ($N = 10$) limits its generalizability, warranting further validation in larger, multi-center cohorts. These findings have translational potential, with microbial signatures like *Alistipes* and *Ruminococcus gnavus* serving as potential biomarkers, while microbiome-targeted interventions such as probiotics or fecal microbiota transplantation (FMT) could help restore beneficial taxa. Moving forward, precision medicine approaches incorporating microbial profiles may optimize treatment strategies, while functional studies are needed to clarify whether dysbiosis is a cause or consequence of CD pathology.

Chapter 5

Conclusion and Recommendations

5.1 Conclusion

This study reinforces the central role of gut microbiome dysbiosis in Crohn's Disease (CD) pathogenesis. Advanced metagenomic and multivariable association analyses identified clear microbial signatures differentiating CD patients from healthy individuals.

Key findings include a notable enrichment of pro-inflammatory taxa such as *Alistipes indistinctus*, *Prevotella copri*, and *Fusobacterium nucleatum* in CD, alongside a significant depletion of beneficial microbes like *Faecalibacterium prausnitzii* and *Roseburia intestinalis*.

The gut microbiome profile also varied with treatment strategies, where immunosuppressants and steroids induced measurable shifts in microbial composition, while mesalamine therapy appeared to promote beneficial species like *Phocaeicola dorei*.

Furthermore, age-related dynamics revealed a decline in important gut commensals over time, which may contribute to disease progression or variability in clinical outcomes.

Despite the robust methodology and statistical rigor, the study acknowledges its limitations, particularly the small sample size, highlighting the need for validation across larger, more diverse populations.

Overall, the results underline the importance of the gut microbiome as both a marker and potential modulator of Crohn's Disease, paving the way for microbiota-informed diagnostic and therapeutic innovations.

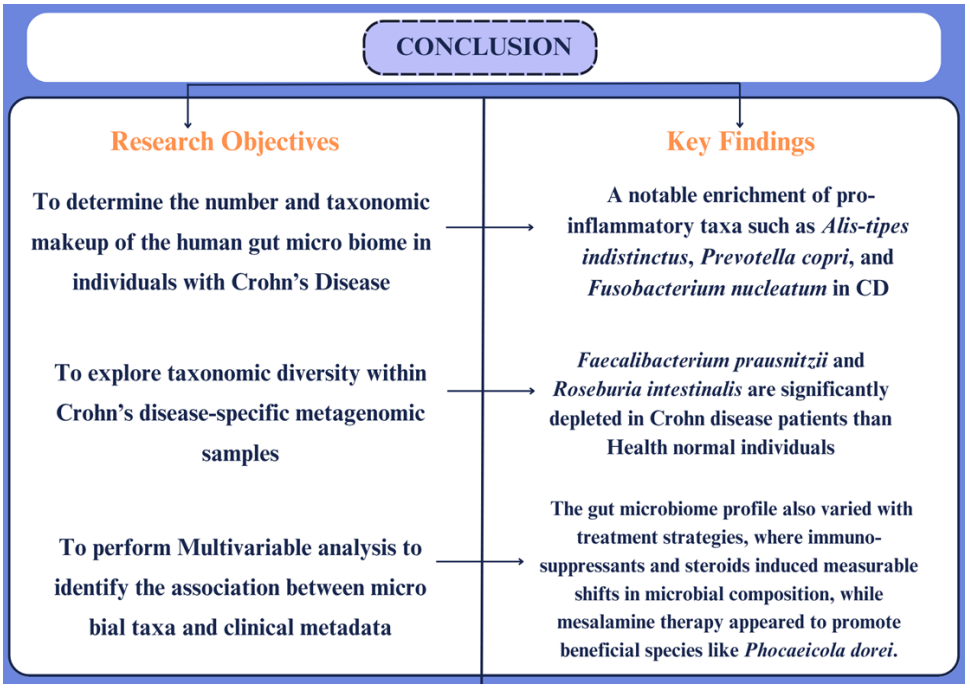

FIGURE 5.1: Conclusion and achievement of research objectives

5.2 Future Recommendations

Future research into Crohn's disease (CD) must prioritize the expansion of study cohorts by incorporating larger, multi-center populations. This approach would enhance the statistical strength of findings and provide a more comprehensive validation of the taxonomic biomarkers identified in current studies. Alongside cohort expansion, there is a pressing need to integrate various functional multi-omics approaches—such as metagenomics, metatranscriptomics, metaproteomics, and metabolomics. By doing so, researchers can achieve a more nuanced and functional understanding of the microbial alterations that occur in CD, offering insights beyond mere compositional shifts.

Equally important is the incorporation of longitudinal study designs that track microbial dynamics over time. Observing the gut microbiome throughout different stages of disease progression—such as flares, remission, and treatment responses—can reveal critical patterns that cross-sectional analyses often miss. In parallel, the concept of personalized microbiome therapeutics should be further explored. Tailored interventions, including specific probiotics, prebiotics, or precision fecal microbiota transplantation (FMT), could significantly improve patient outcomes by aligning treatments with individual microbial profiles.

To uncover causal relationships rather than just associations, mechanistic studies using experimental models like germ-free mice are essential. These models can help determine whether specific microbial imbalances actively contribute to intestinal inflammation in CD. Beyond the lab, there's a growing need for public health initiatives that educate communities—especially in newly industrializing nations like Pakistan—about modifiable risk factors such as smoking and poor diet, both of which can increase susceptibility to CD.

Finally, advancing microbiome-based diagnostics could revolutionize how CD is detected and managed. Developing rapid, non-invasive testing methods based on microbial biomarkers would not only facilitate early diagnosis but also aid in tailoring treatments more effectively, ultimately leading to better long-term management of the disease.

Bibliography

[1] S. Hu, E. Png, M. Gowans, D. E. Ong, P. F. de Sessions, J. Song, and N. Nagarajan, "Ectopic gut colonization: a metagenomic study of the oral and gut microbiome in crohn's disease," *Gut Pathogens*, vol. 13, pp. 1–13, 2021.

[2] R. H. Mills, Y. Vázquez-Baeza, Q. Zhu, L. Jiang, J. Gaffney, G. Humphrey, L. Smarr, R. Knight, and D. J. Gonzalez, "Evaluating metagenomic prediction of the metaproteome in a 4.5-year study of a patient with crohn's disease," *Msystems*, vol. 4, no. 1, pp. 10–1128, 2019.

[3] D.-Y. Kang, J.-L. Park, M.-K. Yeo, S.-B. Kang, J.-M. Kim, J. S. Kim, and S.-Y. Kim, "Diagnosis of crohn's disease and ulcerative colitis using the microbiome," *BMC microbiology*, vol. 23, no. 1, p. 336, 2023.

[4] B. Khorsand, H. Asadzadeh Aghdaei, E. Nazemalhosseini-Mojarad, B. Nadalian, B. Nadalian, and H. Houri, "Overrepresentation of enterobacteriaceae and escherichia coli is the major gut microbiome signature in crohn's disease and ulcerative colitis; a comprehensive metagenomic analysis of ibdmdb datasets," *Frontiers in cellular and infection microbiology*, vol. 12, p. 1015890, 2022.

[5] N. Ding, J. McDonald, A. Perdones-Montero, D. N. Rees, S. Adegbola, R. Misra, P. Hendy, L. Penez, J. Marchesi, E. Holmes *et al.*, "Metabonomics and the gut microbiome associated with primary response to anti-tnf therapy in crohn's disease," *Journal of Crohn's and Colitis*, vol. 14, no. 8, pp. 1090–1102, 2020.

[6] S. Hu, J. Mok, M. Gowans, D. E. Ong, J. L. Hartono, and J. W. J. Lee, "Oral microbiome of crohn's disease patients with and without oral manifestations," *Journal of Crohn's and Colitis*, vol. 16, no. 10, pp. 1628–1636, 2022.

[7] C. R. Armour, S. Nayfach, K. S. Pollard, and T. J. Sharpton, "A metagenomic meta-analysis reveals functional signatures of health and disease in the human gut microbiome," *MSystems*, vol. 4, no. 4, pp. 10–1128, 2019.

[8] O. V. Yvellez, V. Rai, P. H. Sossenheimer, J. Hart, J. R. Turner, C. Weber, K. El Jurdi, and D. T. Rubin, "Cumulative histologic inflammation predicts colorectal neoplasia in ulcerative colitis: a validation study," *Inflammatory bowel diseases*, vol. 27, no. 2, pp. 203–206, 2021.

[9] R. Gowen, A. Gamal, L. Di Martino, T. S. McCormick, and M. A. Ghannoum, "Modulating the microbiome for crohn's disease treatment," *Gastroenterology*, vol. 164, no. 5, pp. 828–840, 2023.

[10] L. J. Cohen, J. H. Cho, D. Gevers, and H. Chu, "Genetic factors and the intestinal microbiome guide development of microbe-based therapies for inflammatory bowel diseases," *Gastroenterology*, vol. 156, no. 8, pp. 2174–2189, 2019.

[11] A. Acharjee, U. Singh, S. P. Choudhury, and G. V. Gkoutos, "The diagnostic potential and barriers of microbiome based therapeutics," *Diagnosis*, vol. 9, no. 4, pp. 411–420, 2022.

[12] N. C. Knox, J. D. Forbes, G. Van Domselaar, and C. N. Bernstein, "The gut microbiome as a target for ibd treatment: are we there yet?" *Current treatment options in gastroenterology*, vol. 17, pp. 115–126, 2019.

[13] M. T. Sorbara and E. G. Pamer, "Microbiome-based therapeutics," *Nature Reviews Microbiology*, vol. 20, no. 6, pp. 365–380, 2022.

[14] Y. Shan, M. Lee, and E. B. Chang, "The gut microbiome and inflammatory bowel diseases," *Annual review of medicine*, vol. 73, no. 1, pp. 455–468, 2022.

[15] M. Rudiansyah, S. Abdalkareem Jasim, B. S. Azizov, V. Samusenkov, W. Kamal Abdelbasset, G. Yasin, H. J. Mohammad, M. A. Jawad, T. Mahmudiono,

S. R. Hosseini-Fard *et al.*, "The emerging microbiome-based approaches to ibd therapy: From scfas to urolithin a," *Journal of Digestive Diseases*, vol. 23, no. 8-9, pp. 412–434, 2022.

[16] Q. Su, Q. Liu, R. I. Lau, J. Zhang, Z. Xu, Y. K. Yeoh, T. W. Leung, W. Tang, L. Zhang, J. Q. Liang *et al.*, "Faecal microbiome-based machine learning for multi-class disease diagnosis," *Nature Communications*, vol. 13, no. 1, p. 6818, 2022.

[17] K. A. Hu and J. Gubatan, "Gut microbiome–based therapeutics in inflammatory bowel disease," *Clinical and Translational Discovery*, vol. 3, no. 2, p. e182, 2023.

[18] L. E. Del Vecchio, M. Fiorani, E. Tohumcu, S. Bibbo, S. Porcari, M. C. Mele, M. Pizzoferrato, A. Gasbarrini, G. Cammarota, and G. Ianiro, "Risk factors, diagnosis, and management of clostridioides difficile infection in patients with inflammatory bowel disease," *Microorganisms*, vol. 10, no. 7, p. 1315, 2022.

[19] Y.-Y. Zheng, T.-T. Wu, Z.-Q. Liu, A. Li, Q.-Q. Guo, Y.-Y. Ma, Z.-L. Zhang, Y.-L. Xun, J.-C. Zhang, W.-R. Wang *et al.*, "Gut microbiome-based diagnostic model to predict coronary artery disease," *Journal of agricultural and food chemistry*, vol. 68, no. 11, pp. 3548–3557, 2020.

[20] C. Callewaert, N. Knödlseder, A. Karoglan, M. Güell, and B. Paetzold, "Skin microbiome transplantation and manipulation: Current state of the art," *Computational and Structural Biotechnology Journal*, vol. 19, pp. 624–631, 2021.

[21] C. Mu, Q. Zhao, Q. Zhao, L. Yang, X. Pang, T. Liu, X. Li, B. Wang, S.-Y. Fung, and H. Cao, "Multi-omics in crohn's disease: New insights from inside," *Computational and structural biotechnology journal*, vol. 21, pp. 3054–3072, 2023.

[22] M. Blencowe and X. Yang, "Found in translation—core network preservation across liver diseases and species," *Cell Reports Medicine*, vol. 2, no. 7, 2021.

[23] Crohnsdi7, "Crohn's disease clinical guidance toolkit," [Online; accessed 2025-05-08]. [Online]. Available: https://gastro.org/clinical-guidance/guideline-toolkits/crohns-disease/

[24] M. M. Elmassry, K. Sugihara, P. Chankhamjon, Y. Kim, F. R. Camacho, S. Wang, Y. Sugimoto, S. Chatterjee, L. A. Chen, N. Kamada *et al.*, "A meta-analysis of the gut microbiome in inflammatory bowel disease patients identifies disease-associated small molecules," *Cell Host & Microbe*, vol. 33, no. 2, pp. 218–234, 2025.

[25] S. C. Ng, H. Y. Shi, N. Hamidi, F. E. Underwood, W. Tang, E. I. Benchimol, R. Panaccione, S. Ghosh, J. C. Wu, F. K. Chan *et al.*, "Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies," *The Lancet*, vol. 390, no. 10114, pp. 2769–2778, 2017.

[26] Drlogy, "6 easy and effective tests for crohn's disease diagnosis | drlogy," [Online; accessed 2025-05-08]. [Online]. Available: https://www.drlogy.com/health/crohns-disease-diagnosis

[27] A. N. Ananthakrishnan, "Environmental risk factors for inflammatory bowel disease," *Gastroenterology & hepatology*, vol. 9, no. 6, p. 367, 2013.

[28] M. A. Núñez-Sánchez, S. Melgar, K. O'Donoghue, M. A. Martínez-Sánchez, V. E. Fernández-Ruiz, M. Ferrer-Gómez, A. J. Ruiz-Alcaraz, and B. Ramos-Molina, "Crohn's disease, host–microbiota interactions, and immunonutrition: Dietary strategies targeting gut microbiome as novel therapeutic approaches," *International Journal of Molecular Sciences*, vol. 23, no. 15, p. 8361, 2022.