University of Padova

Department of Physics and Astronomy "Galileo Galilei"

Master Thesis in Physics of Data

Data-Driven Analysis of Microbial Community Dynamics in Inflammatory Bowel Diseases

Supervisor
Prof. Samir Simon Suweis
University of Padova

Co-supervisor
Dr. Sonia Facchin
University of Padova

Master Candidate
Ekaterina Chueva

Student ID
2072050

Academic Year
2023-2024

Words cannot express my gratitude to my family and my dear friend Polina, who supported me through every step of my studies. I am also thankful to all the friends that I was lucky to find here in Italy. You made this journey truly special for me.
With sincere recognition, I express my utmost thanks to my supervisors, prof. Samir Suweis and Dr. Sonia Facchin, for giving me opportunities to learn and grow, as well as for providing guidance and thoughtful feedback.

Abstract

Modern technological advances allow study of microbial communities with unprecedented precision. It is crucial to investigate and understand these communities in disease, as it can lead to discovery of new treatments. This thesis aims to explore gut microbiome dynamics in inflammatory bowel disease, in particular with the influence of the sodium butyrate treatment. To achieve this goal, a dataset of 16S rRNA sequences obtained from stool samples of patients with Crohn's disease and ulcerative colitis was used as well as data from healthy donors for comparison. The patients were participants of the sodium butyrate drug trial, thus the study of its efficiency was possible. The bioinformatics analysis was performed using QIIME 2 software, followed by functional assignment with a use of Genomic Content Network and KEGG annotations. Then, functional composition of the gut microbiome was studied. As a result, previously reported distinguishing patterns for healthy and unhealthy microbiome were observed. It was found that sodium butyrate treatment shows a positive influence on patients with Crohn's disease, which can be seen through function/function Pearson correlation values histogram and eigenvalue decay profile of correlation matrices; while patients with ulcerative colitis did not show any improvements.

Contents

Listing of figures

Listing of tables

Listing of acronyms

IBD . . . . . Inflammatory bowel disease

CD . . . . . Crohn's disease

UC . . . . . Ulcerative colitis

GI . . . . . Gastrointestinal

SFCA . . . . . short-chain fatty acids

OTU . . . . . Operational taxonomic unit

ASV . . . . . Amplicon Sequence Variant

NGS . . . . . Next-generation sequencing

HD . . . . . Healthy donors

TD . . . . . Taxonomic dataset

GCN . . . . . Genomic content network

FD . . . . . Functional dataset

KS . . . . . Kolmogorov-Smirnov

1 Introduction

The theoretical investigation of biological systems has been attracting attentions of physicists since last century. A father of modern computing, Alan Turing, proposed reaction-diffusion theory in his work "The Chemical Basis of Morphogenesis" (1952) in order to explain formation of spatial patterns like the ones on animal skins. Simple reaction equations can be also used for describing self-organization of micelles. This description plays a crucial role for understanding biological processes as, for example, membrane formation. More sophisticated methods were applied for biological systems as well, such as statistical mechanics: for example, it was used to model genetic mutations (Max Delbrück), interactions and movement of molecules in cells (via random walk model), modelling of mechanical properties of polymers (e.g. wormlike chain model for describing flexible and semi-flexible polymers). A Nobel Prize in Physics (2024) was given to John Hopfield, who proposed energy-minimizing systems, analogous to a spin glass model, that represent associative memory. This work in particular bridged the gap between physics and neuroscience, inspiring modern AI research.

Physicists contributed not only in theoretical research of biological systems, but also in experiments. Rosalind Franklin, biophysicist and X-ray crystallographer, obtained diffraction images of DNA (1953), that unrevealed its double-helix structure. Such modern technologies, as optical tweezers, allow to detect molecular motors (e.g. kinesin) or to explore DNA's mechanical properties. Besides that, experiments of Alan Hodkin and Andrew Huxley on neural signaling (1952) laid a foundation for modern computational

neuroscience.

Modern advances in technologies, such as next-generation sequencing (NGS), disclose the possibility of investigating a particular type of biological systems, microbial communities, with unprecedented resolution. Nowadays, NGS experimental data has been made publicly available to researches on such resources as NCBI Sequence Read Archive (SRA), EMBL-EBI European Nucleotide Archive (ENA), Gene Expression Omnibus (GEO) and many others. This "data explosion" is both challenging, as it requires inventing new tools for analysis, but also an opportunity for many researchers to identify rules driving the functioning of biological systems.

This thesis aims to explore microbial community's dynamics in gut in inflammatory bowel disease, using data-driven approaches. Besides that, influence of the sodium butyrate treatment is studied as well.

First, in this chapter an overview of existing approaches of studying gut microbiome is discussed as well as why is it important for inflammatory bowel diseases in particular. In chapter 2 the dataset used in this work is defined. Chapter 3 is dedicated to the pipeline for bioinformatics analysis and functional inference. Finally, in chapter 4 the results are reported.

1.1 Inflammatory bowel disease and gut microbiome

Inflammatory bowel disease (IBD) is a chronic disease of gastroenteric tissue, which leads to episodes of inflammation [1]. IBD has two disorders – Crohn's disease (CD) and ulcerative colitis (UC). During CD, areas of inflammation alternate with normal-appearing mucosa anywhere in the gastrointestinal tract, while UC affects only the colon, resulting in continuous inflammation [2].

IBD has become a global disease. Since 1990, while in such regions as North America, Oceania and Europe this disease mainly shows stable or decreasing incidence, in newly industrialized countries in Africa, Asia and South America diagnosis of IBD has been growing [3]. However, even in western countries IBD affects 1 in 200 individuals [3].

The cause of the IBD appears to be related to abnormal host immune responses to the intestinal microbiota [1]. Therefore, before embarking upon microbiome in IBD, first the thorough description of the microbiome and its influence on human health should be provided.

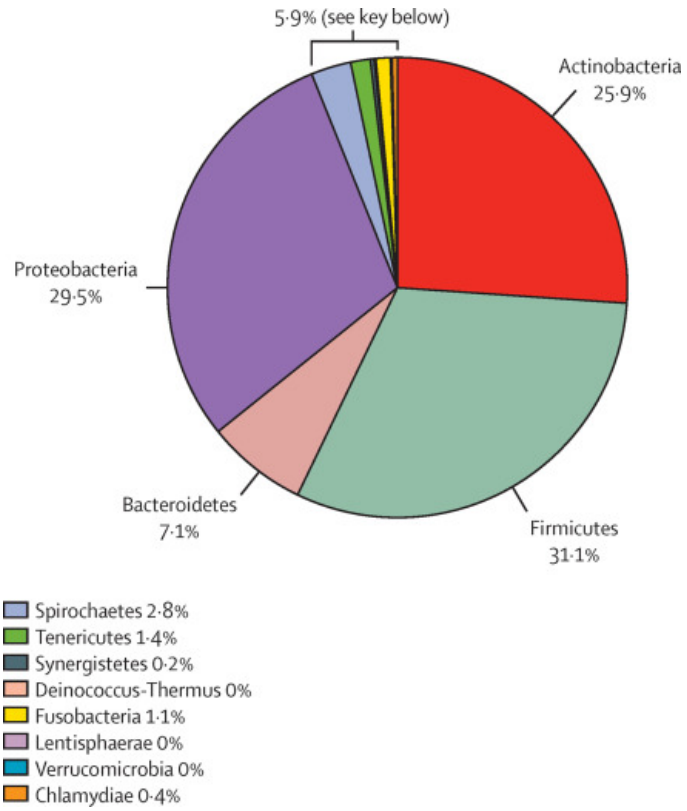

Figure 1.1: Phylum-level taxonomic composition of the human gut microbiome [7].

The human gut is home to bacteria, viruses, fungi and protozoa, resulting in 100 trillion different microbial organisms [4]. In fact, the gastrointestinal (GI) tract is one of the largest borders between environmental factors, antigens and host in a human body [5]. The composition of these microbial organisms in GI tract is referred to as gut microbiome or microbiota. In healthy microbiome, cellular relative abundances is much greater for bacteria [6]. In more detail, it has been found with a combined dataset [7, 8] of MetaHit [9] and Human Microbiome project [7] that 93.5 % of the phyla present in human gut belongs to Proteobacteria, Firmicutes, Actinobacteria and Bacteroidetes (fig.1.1). Due to the environmental conditions of the colon, the majority of the bacteria are anaerobes. However, there exist gradients of oxygen across the colonic wall, thus some of bacteria are facultative anaerobes [10], meaning that their energy cycles utilize oxygen if it is available and is anaerobic if it is not available.

Gut microbiome's symbiotic interaction with a host (human) is essential to the human health, since it is related to the supply of nutrients and energy, development of immune

system and host defense against pathogens [4].

One of the important processes in gut is bacteria expression of carbohydrate-active enzymes, that later generate metabolites such as short-chain fatty acids (SFCAs). These SFCAs can be absorbed by colon, playing a role in gut motility, inflammation, glucose homeostasis [11]. The most abundant SFCAs are acetate, propionate and butyrate (with a proportion of 3:1:1), which together have a concentration of 50-150 mM in colon [12]. In particular, butyrate is known for its anti-inflammatory and anticancer effects, and is an important energy source colonocytes [5]. Among other SFCAs influence on host's health are modulating appetite regulation and energy intake.

Another key activity of the gut microbiome is the *de novo* synthesis and supply of vitamins, as humans lack biosynthetic capacity for most vitamins and thus these should be provided exogenously [13]. Gut microbiome provides vitamins K and B (B_2 and B_{12}), biotin, folates and possibly others [11]. For example, vitamin B_{12} is mainly synthesized by lactic acid bacteria and it cannot be produced by animals, plants or funghi [5].

As for the immune system, it was found that gut microbiome is important for both intestinal mucosal and systemic immune system [5].

Lastly, another important function of the microbiome is protection from pathogens [14]. Microbiome influences pathogen colonization using activities such as, for example, competing for attachment sites or nutrient sources [5].

The above important role of the gut microbiota is described in health. Dysbiosis, a disadvantageous change in gut microbiota, alternates the host-microbiota interaction. Dysbiosis in IBD is characterized by reduced diversity of the microbiota, decrease of bacteria that produce short chain fatty acids, increase of mucolytic bacteria and increase in pathogenic bacteria [4]. IBD has been long associated with dysbiosis, as composition and diversity of the gut microbiota is a key factor of developing IBD [15]. However, no definitive cause-effect relationship between them has been proven for humans [16]. Thus dysbiosis is often referred to be as a "feature" of the IBD.

Nowadays, growing attention is drawn to potential treatments for IBD by influencing gut microbiome [17]. Thus, a thorough understanding of the gut microbiome is needed, in order to characterize better its composition and function in IBD patients. Additionally, studying gut microbiome before and after treatment is needed for the evaluation of the treatment in question.

1.2 16S rRNA sequences for studying gut microbiome

16S ribosomal RNA marker gene has become widely used to access microbiome diversity. The choice of 16S rRNA gene is dictated by several of its functions, such as ubiquity and evolutionary properties [18]. In more detail, as reported in [19], 16S rRNA is present in almost all bacteria, its function over time has not changed and it is big enough for bioinformatics purposes. 16S rRNA gene plays a crucial role in cellular function and survival [20]. The majority of the 16S rRNA-based genotype protocols use $V_5 - V_6, V_3 - V_4$ or V_4 hypervariable regions out of the nine variable regions ($V_1 - V_9$), which with a highly conserved primer binding sites constitute the gene sequence [20]. The alternation of highly conserved and hypervariable regions allow researchers understand both phylogenetic relationship of distant organisms as well as differentiate organisms on genus- and family-levels.

In order to obtain 16S rRNA sequences for microbiome analysis, it is needed to perform stool sample collection. There are three key parameters to address during sample collection and handling [20]: contamination, transportation, storage and safety. The first aspect is important due to different sample environment or proximity of other samples leading to unreliable results. The second aspect refers to microbial community instability, thus immediate freezing and maintaining the same conditions are principal. The last aspect underlines the importance of maintaining of these conditions before sequencing.

The pipeline for sequencing 16S rRNA in next or third-generation sequencing methods includes several steps: genomic DNA extraction, targeted PCR amplification, library construction and finally sequencing [21]. More specifically, Polymerase Chain Reaction (PCR) amplification stage targets only specific regions mentioned above; and a library construction step involves collecting DNA fragments from samples that have special adapters added to both of their ends. Then, sequencing is performed from one end to another of each fragment and each base is detected by a unique signal that is specific to it. When the signal is sufficient to determine the base, it is written along with a quality score that quantifies the confidence in the base call [22]. As a result, a FASTQ files are obtained. These files are text files that shows amino acid sequences with quality score.

After the described above procedure it is possible to use a specific software for 16S sequence analysis with further taxonomic identification.

1.3 Bioinformatics analysis of the microbiome

The bioinformatics pipeline is needed for microbial identification. In microbiome research, methods are using either Operational Taxonomic Units (OTUs) or Amplicon Sequence Variants (ASVs). The general pipeline of bioinformatics analysis consists of working with raw sequences (quality-filtering and handling errors in sequences), obtaining some form of sequence representations (OTUs or ASVs) and taxonomic classification. The details for both methods are provided below.

The first group of methods, named clustering methods, use OTUs. OTUs are clusters that are grouped together based on a sequence identity above a chosen threshold, which is usually set to 97% [23]. This clustering helps to reduce the size of the dataset and the influence of sequences with errors. The pipeline for such methods involve primer trimming, quality-filtering, sequence clustering, chimera removal, generating OTU table and taxonomic classification [24]. A popular tool for OTU-based sequence analysis is MOTHUR [25].

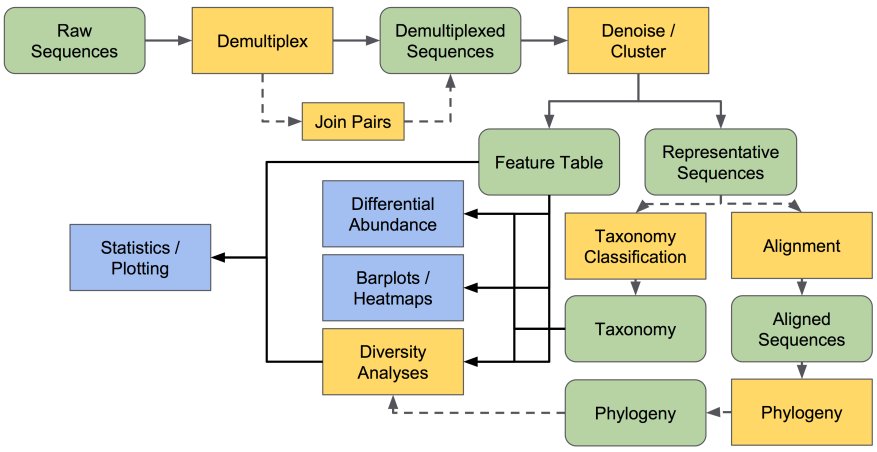

Figure 1.2: Bioinformatics pipeline example, flowchart for ASV data in QIIME 2 [26].

The second, alternative, group of methods (ASV-based) is called denoising methods. These methods provide higher resolution, as they identify exact sequence after error correction, without clustering. The general pipeline for such methods include quality filtering, error correction, chimera removal, generating ASV table, taxonomic classification. A popular tool for implementing ASV-based method is DADA2 [27]. Besides that, in QI-

IME 2 [28] software both OTU- and ASV-based methods are provided, leaving a choice to the user.

For both groups of methods the last step of the analysis is taxonomic classification. This classification is done through machine learning methods such as, for example, Naive Bayes classifier, or mapping to reference databases, for example SILVA [29] or Greengenes [30]. The mentioned above tools (MOTHUR, DADA2 and QIIME 2) have built-in functions in order to perform taxonomic classification. As an example, in fig. 1.2 the flowchart of a typical bioinformatics pipeline for the ASV data performed in QIIME 2 is shown.

1.4 Functional analysis of the gut microbiome

After performing taxonomic classification, a logical subsequent step would be inferring functions from taxonomy in order to further analyze microbial community. However, it is not possible to achieve directly from 16S rRNA sequences and additional software or databases are required.

There are several widely used databases for classification of bacterial functions. Among them there are:

- KEGG (Kyoto Encyclopedia of Genes and Genomes) [31]: a database that classifies functions based on pathways, ortholog groups and metabolic networks;
- eggNOG (evolutionary genealogy of genes: Non-supervised Orthologous Groups) [32]: extended orthologous group classification with functional annotations;
- COG (Clusters of Orthologous groups) [33]: classification focusing on grouping proteins into families with similar functional roles;
- CAZy (Carbohydrate-Active enZymes Database) [34]: useful database for studying bacterial carbohydrate metabolism;
- MetaCyc [35]: a database which focus is on describing metabolic pathways and enzymes;
- Pfam (Protein Families Database) [36]: classification of protein domains and families.

These databases provide reference in order to assign functional roles based on taxonomic data. There are around ≈ 20 tools that perform such assignment [37]. The most

popular tools are PICRUST2 [38] and Tax4Fun [39]. However, it is possible to use directly the mentioned above databases for functional annotation.

2

Dataset

This thesis aims to explore dysbiosis in IBD patients and the efficiency of the sodium butyrate treatment. In order to do this, a comprehensive dataset was chosen. This section outlines key aspects and details of the dataset source, experimental setting, and its characteristics.

The dataset consists of IBD patients and healthy donors (HD), the latter 24 in total. The details for the IBD patients are provided below. Sequencing was performed similarly for both groups.

2.1 Experimental setting

The study of sodium butyrate treatment was conducted at Azienda Ospedale Università of Padua (Italy) from September 2020 to June 2022. Sodium butyrate drugs and placebo were provided by SILA srl, Noale Venice, Italy. Drug and placebo pills had similar appearance and taste. The study was funded by Department of Surgery, Oncology, and Gastroenterology, University of Padua, and SILA srl, Noale. This study was conducted by S.Facchin, M.Calgaro, M.Pandolfo, A.Buda, B. Barberio, F. Zingone, N. Vitulo, E. V. Savarino in order to evaluate sodium butyrate treatment efficiency [40, 41]. S.Facchin provided the raw data for an alternative analysis described in this thesis.

The choice of sodium butyrate was dictated by the fact that SFCAs, as was described

before, are among the crucial types of microbiome bio-products. A presence of butyrate in colon depends on the amount of butyrate-producing bacteria, cross-feeding interactions with the gut microbiome and fibers consumed. Butyrate is the main energy source for coloncytes, which uses of 95 % of all colonic butyrate. Besides that, it is crucial for the intestinal immune system. Previous studies suggested that butyrate treatment could be beneficial for IBD patients, however it was not clear for which type of IBD.

The trial was randomized placebo-controlled and prospective. The patients involved in the study had a histologically confirmed diagnosis of IBD in the last 6 months, within an age range 18-80. IBD patients who agreed to participate and who met the conditions for participation in the study were randomly assigned to a drug or a placebo with probability 50 %. The drug or placebo was taken by IBD patients 3 capsules per day for 90 days.

The assignment to interventions was done by a non-involved in a study nurse and concealed, thus all further processing of the data was performed blindly. Since, as stated before, the pills had similar taste and appearance, the trial was double-blind.

Besides sodium butyrate or placebo, the patients were receiving a conservative therapy for IBD.

2.2 Dataset details

Overall, there were 140 IBD patients in the age range 18-80 years old in total. Among them, 60 were diagnosed with CD and 80 with UC. Since the metadata for these patients was also provided, it was possible to make descriptive plots shown on fig.2.1. Stool samples were collected before and after 90 days of the treatment.

Collected samples were solubilized, stabilized and then stored at -20°C. The $V_3 - V_4$ regions of the 16S rRNA data were amplified with the use of these two primers:

- Pro341F: 5'-CCTACGGGNBGCASCAG-3';
- Pro805R: Rev 5'-GACTACNVGGGTATCTAATCC-3'.

The modification of primers, needed for dual-index library preparation, was done with forward overhang: 5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG [locus-specific sequence]-3' and with reverse overhang:
5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG [locus-specific sequence]-3'.

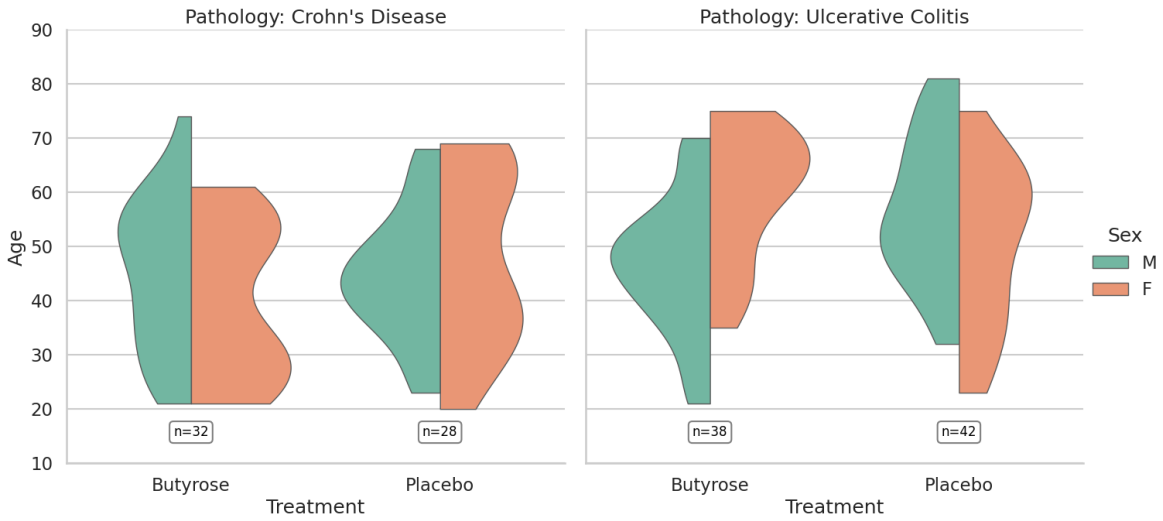

Figure 2.1: Descriptive plots for IBD patients: distribution of the patients' ages for both males and females; the total number of patients receiving either sodium butyrate or placebo specified below the plots.

This procedure was done following Illumina protocol. After this, samples were normalized, pooled and run on Illumina MiSeq with 2×300 bp approach.

The dataset is in the form of FASTQ files. In total, there are 280 files for IBD patients, as for each patient there were two time points for stool sample collection, before and after treatment. Each FASTQ file contains raw sequencing data, which includes these key components for each DNA read:

- sequence identifier;
- nucleotide sequence;
- separator lines;
- quality scores.

From the metadata file, the following entries were used in this thesis:

- pathology (CD or UC);
- treatment (sodium butyrate or placebo);
- time point (T_0 – before treatment, T_1 – after 90 days of treatment).

The FASTQ files were then processed using bioniformatics software, followed by functional assignment. These two procedures are described in detail in the next chapter.

3 Methods

This chapter is dedicated to the bioinformatics analysis and further functional analysis of the microbiome. First, the bioinformatics pipeline is discussed. The result of this step is obtaining taxonomy and counts of the gut bacteria found in the samples. Second, the methodology for mapping of functions to the taxonomy is examined in details.

3.1 Bioinformatics analysis

The bioinformatics analysis was done with QIIME 2 2024.2 [28]. Overall, there were three main steps in this analysis:

1. preparing raw sequences;
2. denoising;
3. taxonomy assignment.

Preparing raw sequences started with importing FASTQ files into QIIME 2 and demultiplexing them with *q2-demux* plugin. At this step, using the visualization tools available in QIIME 2, the number of sequences found in each sample was explored. An example of this step's visualization is shown in the fig.3.1. Undersampling was discovered for one sample, thus it was discarded. Then, the quality plots were explored in order to determine

truncating lengths for forward and reverse reads. An example of the interactive quality plot is shown in the fig.3.2. Based on these plots, the truncation lengths, used later in the pipeline, were set to:

- 280 (forward reads) and 220 (reverse reads) for IBD patients;
- 280 (forward reads) and 205 (reverse reads) for HD.

The last procedure at preparing raw sequences was removing primers using *q2-cutadapt* plugin within QIIME 2 [42]. This procedure was done separately for the batches of data both for IBD patients and HD.

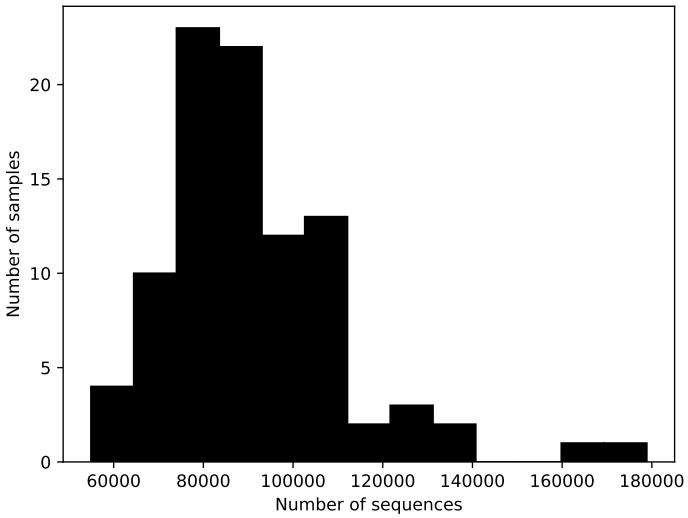

Figure 3.1: Example of a distribution of number of sequences found in each sample for one batch of data for IBD patients (forward reads), QIIME 2 visualization.

The second step of the bioinformatics analysis was denoising. Denoising was performed using DADA 2 [27] implementation in QIIME 2. At the same time, the sequences were truncated with the specified above lengths as parameters. This procedure was done for batches of data as before. As the result, the ASV tables were obtained for each batch of data. These tables track the number of times a given ASV was observed in each sample. Finally, using *feature-table merge* function, all ASV tables were combined into two tables, which correspond to IBD patiens and HD. Following QIIME 2 notation, these tables are called “feature tables”. An example of a feature table is shown in the fig.3.3. Overall, 5134 and 1082 features were found for IBD patients and HD respectively.

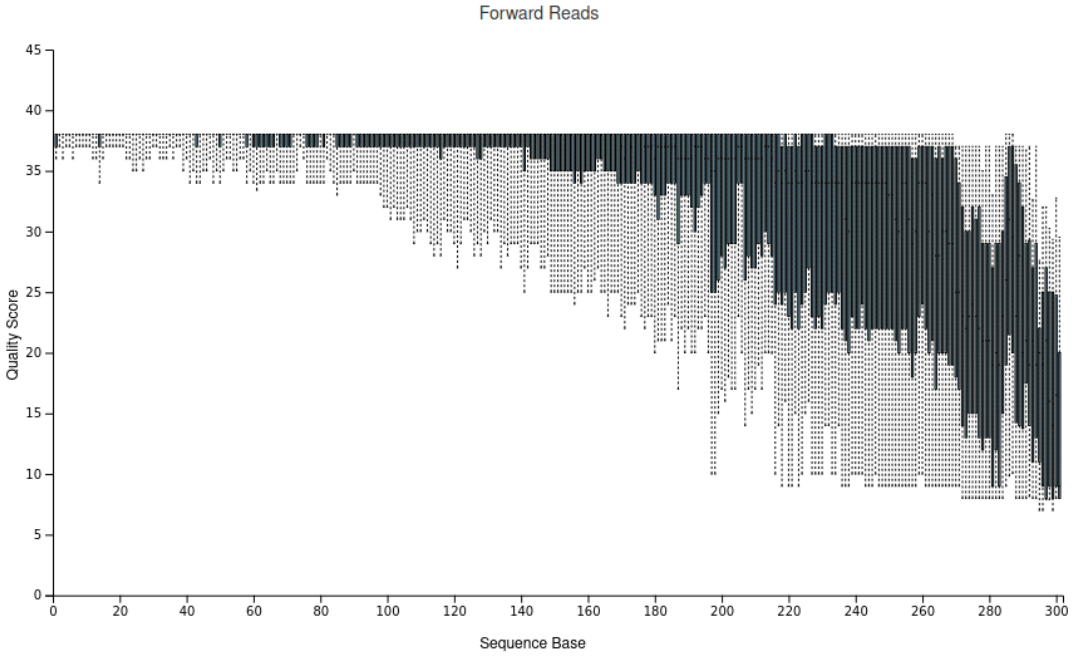

Figure 3.2: Example of an interactive quality plot in QIIME 2 for one batch of data of IBD patients (forward reads).

Last step is the taxonomy assignment. The SILVA 138.1 [29] was chosen as a reference database. In order to make a QIIME 2-compatible database, based on NR99 SILVA, RESCRIPt was used [43]. For the purpose of making a reference database for the specific dataset, the low-quality sequences were culled (sequences containing 5 or more degenerate pairs) and the reads for amplicon region in interest were extracted, using the primers specified above. After these steps, a ready-to-use sequence and taxonomy files were saved. The taxonomy assignment was performed to the ASV tables through *q2-feature-classifier* plugin [44] with Naive Bayes classifier [45]. This step was done for IBD patients and HD

	Frequency	# of Samples Observed In
4abaa483334092f021534a979086baeb	761,459	155
b6635d67cb594473ddba9f8cfba5d13d	584,176	172
6251bd9ebf43fae466939ab366f6e547	354,781	150
22f4ee9a41a4d73580bf7ade8e9e017a	334,872	223
70d55baf78e9ac4d0babeac5dcbae5c2	282,421	175

Figure 3.3: Example of the ASV table: first five entries of the ASV table for IBD patients, QIIME 2 visualization.

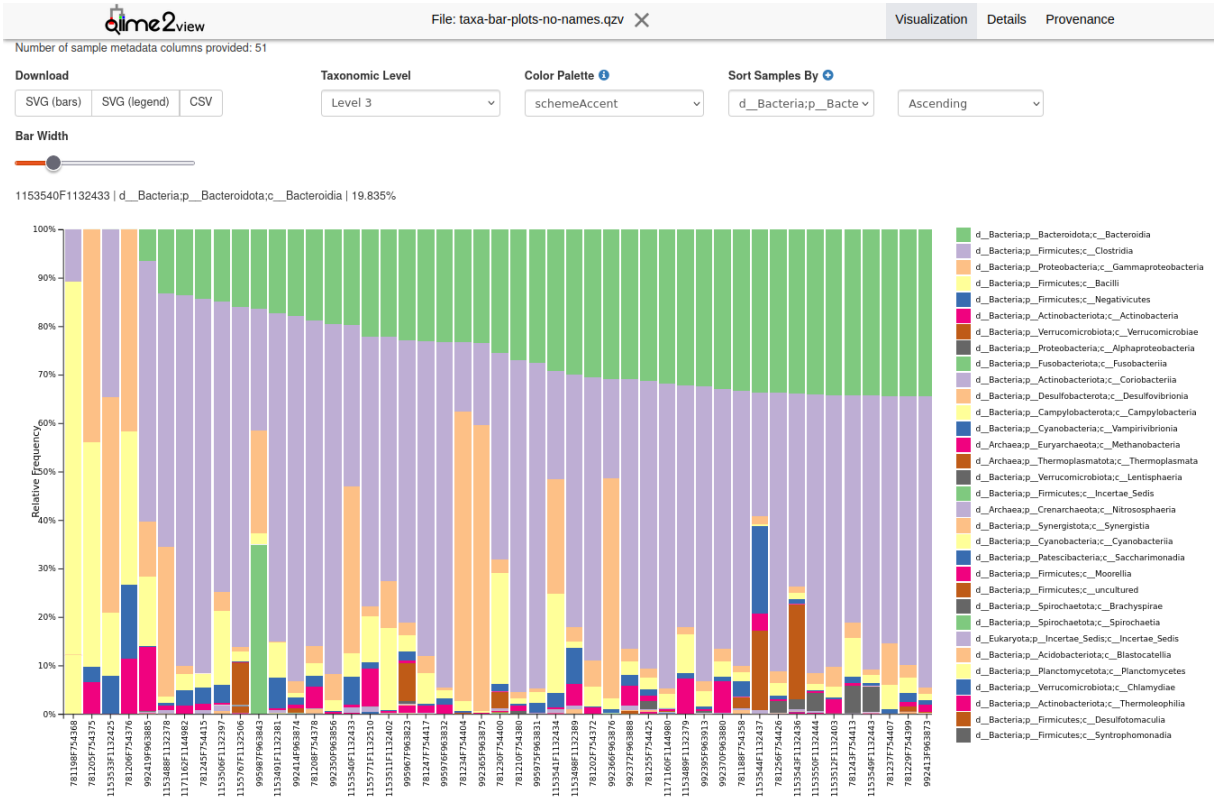

Figure 3.4: Interactive QIIME 2 interface for taxonomic assignment on a class level for IBD patients: sample ID and frequency of the taxa are on the x and y axis respectively, colours define different taxa.

separately.

Final procedure in the bioinformatics analysis was pre-filtering taxonomy and choosing its level. In general, QIIME 2 provides these levels of taxonomy assignment: domain, phylum, class, order, family, genus, species. First, using *taxa filter-table* taxa were pre-filtered, excluding entries that were not assigned to the phylum level. Then, metadata files (for HD the metadata was just names of the samples) were merged with the obtained taxonomy in order to receive the interactive visualization for taxonomic assignment (fig.3.4). Using this interactive interface, the csv files for the genus-level taxonomy were saved. Genus level was chosen due to the fact that in QIIME 2 taxonomic classification done against the SILVA database does not provide species-level classification.

In the end, the output of the bioinformatics pipeline were tables in csv files (separate for IBD patients and HD), in which there were samples' numbers as rows and taxonomy labels as columns.

3.2 Inference of bacteria functions

As discussed before, inferring functions from taxonomy is not possible in bioinformatics software. Obtaining functions was done following methodology developed in [46]. In this section taxonomic datasets (TDs), Genomic Content Network (GCN) and Functional datasets (FDs) are defined as well as procedure of constructing these datasets and network is described.

3.2.1 Taxonomic dataset

First, the TDs are constructed for HD and IBD patients. More precisely, for IBD patients TDs were done for the following groups separately (for each group a further division before and after treatment was considered as well): CD patients who received sodium butyrate treatment, CD patients who received placebo, UC patients who received sodium butyrate treatment and, finally, UC patients who received placebo. Overall, there were constructed eight TDs for IBD patients and one TD for HD.

As stated above, the outputs of QIIME 2 pipeline are tables with samples' numbers as rows and taxonomy up to genus level as columns. The pre-processing of these tables consisted of the following steps:

1. Removing all the missing assignments.

   After this procedure, only valid genus level labels stay in the tables.

2. Combining some of taxonomic assignments.

   Since later the Unified Human Gastrointestinal Genome (UHGG, version 2.0.1) [47] database is used for inferring functions, having the labels that are found in this database is essential. Thus, some of the assignments were combined into existing in this database labels, e.g. *Clostridia_UCG– 014* and *Clostridia_vadinBB60_group* were united into *Clostridia* column.

3. Discarding all the taxonomic labels that were not found in the UHGG database.

After these steps, all the tables were transposed and normalized column-wise, resulting in a vector of relative bacteria abundances for each sample. After this, it is necessary to remove the false-positives appearing in the tables. Thus, the threshold was applied.

Following [46], this threshold is chosen as the logarithmic flex point of the curve of the average (α) local diversity as a function of threshold values:

$$\eta_0 = \underset{\eta}{\operatorname{argmin}} \frac{d\langle\alpha(\eta)\rangle}{d\log(\eta)}. \tag{3.1}$$

Alpha diversity was calculated using *diversity.alpha_diversity* function from *skbio* package [48] with chosen Shannon entropy as a metric. After this, the tables were renormalized again, to maintain the normalization. The resulting tables are called TDs. TDs were made separately for IBD patients and HD.

3.2.2 Genomic content network

Second, Genomic Content Network (GCN) of microbiome was constructed both for IBD patients and HD. GCN is a bipartite graph connecting microbes to the genes in their genomes [49]. GCN provides all the information about functional overlap of different microbes in microbial communities. In other words, GCN provides mapping of bacterial taxa to its functions. Furthermore, GCN's links are weighted, representing the appearance of a given function in a particular genome.

In order to construct GCN, the reference database and function annotations need to be chosen. As mentioned above, UHGG (version 2.0.1) [47] was selected as a reference database. The functional annotation adopted in this work is KEGG [31], in particular KEGG pathways.

In detail, GCN (separately for IBD patients and HD) were constructed following this pipeline:

1. Finding genus in interest in the database.

   For each taxa from TDs, the corresponding entries (species level) were found in the database. For each genus label, a list of links to these entries were collected.

2. Finding counts of each KEGG pathway for each genus.

   Using the previous lists of corresponding to each genus entries, the total counts of each KEGG pathway for each genus were obtained.

3. Normalization.

   First, each KEGG pathway count was divided by the number of species found in the database for each genus. Then, for each taxa the counts were normalized to one.

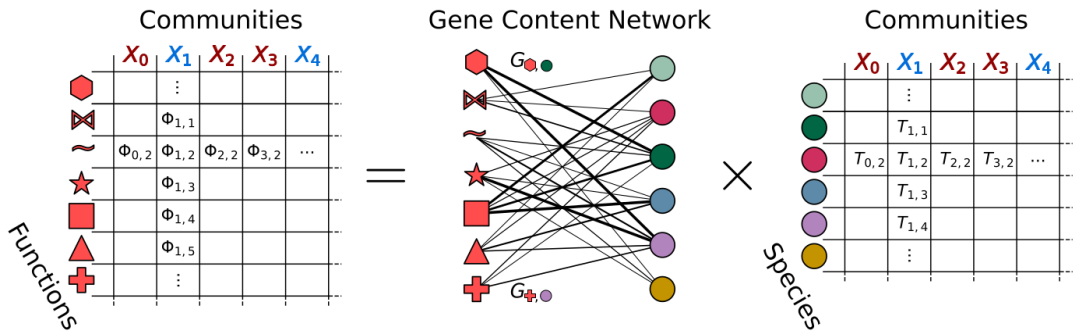

Figure 3.5: Visual representation of construction of functional dateset from taxonomic dataset and genomic content network [46].

The result of the previous steps is two GCNs, for IBD patients and HD, that provide average functional profiles within each bacteria genus.

3.2.3 Functional dataset

The last step needed for further functional analysis is obtaining FDs both for IBD patients and HD. Conceptually, this step is shown in the fig. 3.5.

If we define FD as Φ, TD as T and GCN as G, then the equation for FD would be [46]:

$$\Phi_{fs} = \sum_{t \in \text{taxa}} G_{ft} T_{ts}. \tag{3.2}$$

Here indexes f,t, and s correspond to functions, taxa, and samples respectively.

FDs were constructed for the groups specified above, thus eight FDs were made for IBD patients and one FD was done for HD.

4 Results

In this chapter, the obtained functional datasets for HD and IBD patients are analyzed in order to determine the functional patterns found in healthy and unhealthy microbiome and the influence of the sodium butyrate treatment on CD/UC patients.

4.1 Functional analysis

In the previous chapter, the procedure for making functional datasets was discussed. As stated above, there were eight functional datasets for IBD patients (subsetting data was done based on such criteria as UC/CD, treatment/placebo and before/after treatment time points) and one functional dataset for HD. However, before doing this, the taxonomic datasets were also explored in order to verify that they are not representative for the analysis, as was discussed in [46].

First, the Pearson taxa/taxa and function/function correlation matrices were calculated. The elements of these matrices are equal to:

$$\rho[i,j] = \frac{\sum_{k=1}^{m}(X_{ki} - \overline{X}_i)(X_{kj} - \overline{X}_j)}{\sqrt{\sum_{k=1}^{m}(X_{ki} - \overline{X}_i)^2}\sqrt{\sum_{k=1}^{m}(X_{kj} - \overline{X}_j)^2}},$$

where X represents either a taxa or a function, $\overline{X}_i = \frac{1}{m}\sum_{k=1}^{m} X_{ki}$ – mean of the i-th sample

over all taxa/functions. It was found that, possibly due to the nature of the datasets, correlation matrices were not positive semidefinite. In order to fix it, the algorithm for finding the nearest positive semidefinite matrix was applied, based on [50]:

Algorithm 4.1 Nearest Positive Semidefinite Matrix (Higham's Method)

Require: $M \in \mathbb{R}^{n \times n}$, maximum iterations $k_{\max}$, tolerance δ
Ensure: Nearest positive semidefinite matrix X
Initialize $Y \leftarrow M$
Set weighting matrix $W \leftarrow I_n$ to an identity matrix
for $k = 1$ to $k_{\max}$
    Eigenvalue decomposition: $Y = Q\Lambda Q^\top$
    Replace negative eigenvalues in Λ with zero
    $X \leftarrow Q\Lambda Q^\top$
    $X \leftarrow \frac{1}{2}(X + X^\top)$
    $R \leftarrow Y - X$
    $Y \leftarrow X + WRW$
    Compute Frobenius norm $\|R\|_F$
    if $\|R\|_F < \delta$
        Return X
    end if
end for
Warning: Did not converge within $k_{\max}$ iterations
Return X

For all the datasets the algorithm 4.1 converged with $\delta = 10^{-12}$. Then, the results were summarized into histograms and are shown in fig. 4.1 for TDs and in fig.4.2 for FDs. In order to calculate probability density function shown in these figures, to original histograms with 30 bins the *gaussian_kde* function from SciPy [51] Python package was applied.

As it can be observed from the fig.4.1, it is not representative for the analysis, as there is no significant difference between healthy and unhealthy microbiome taxonomic correlation histograms. Thus, the further analysis was done for function-function correlation matrices.

In order to interpret the results shown in fig. 4.2, it is necessary to mention some key findings of [46]: there exist a core of strongly correlated functions in a healthy microbiome, while in unhealthy one more uncorrelated functions are observed. Since in the dataset used in this thesis HD were provided as well, it was possible to verify this finding.

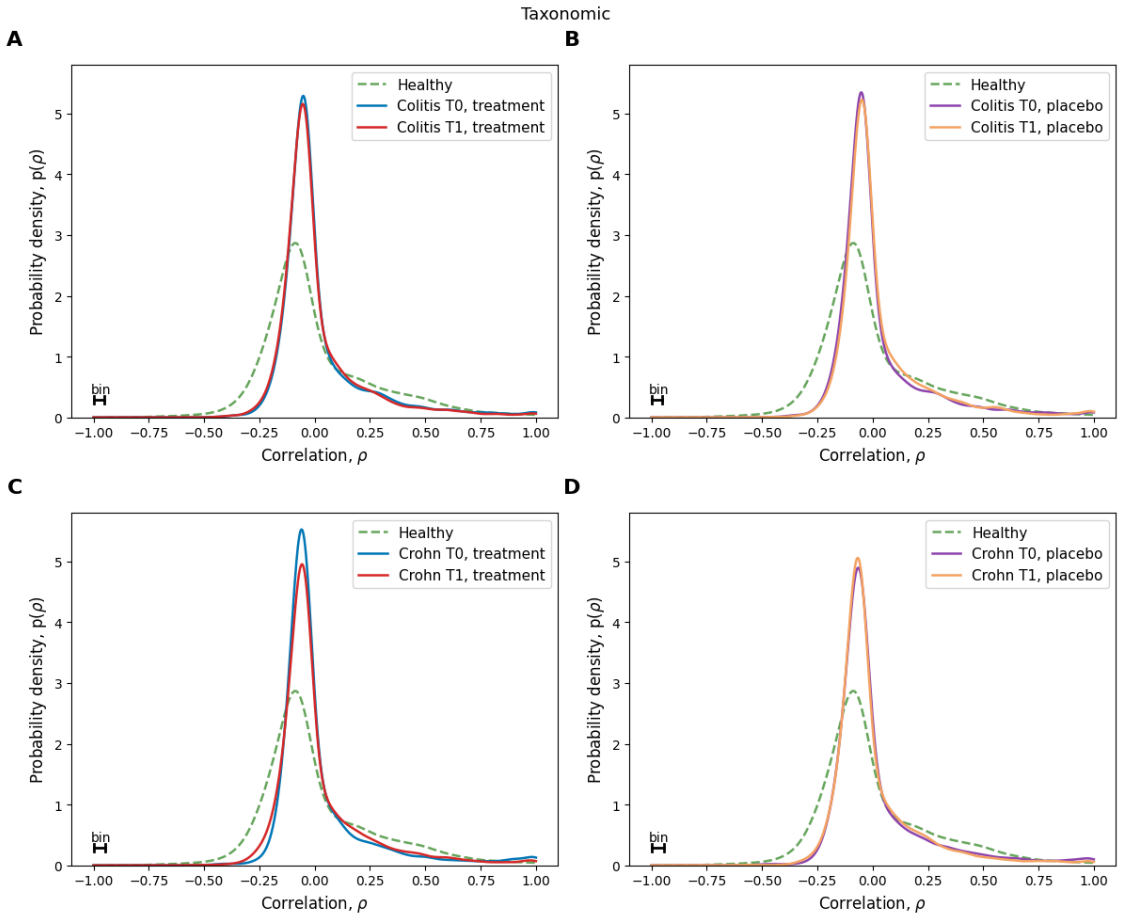

Figure 4.1: Histograms of the values of the taxa/taxa correlation matrix before (T_0) and after (T_1) treatment. HD group put as a reference. (**A**, **B**) UC patients, treatment/placebo. (**C**, **D**) CD patients, treatment/placebo.

Indeed, the histograms values near +1 and -1 in the fig.4.2 are greater for HD than for IBD patients. The further interpretation should include efficiency of the sodium butyrate treatment. If the treatment is efficient, then the probability density functions of 'after treatment' subgroups for IBD patients who received sodium butyrate should be more resembling the healthy one compared to 'before treatment'; and should stay the same for placebo subgroups for both 'before' and 'after' treatment time points. As it can be seen, for CD patients this pattern is observed, while for UC patients there are no observable improvements after receiving sodium butyrate treatment.

To verify the above conclusions, the two sample Kolmogorov-Smirnov (KS) statistical test was performed, using the probability densities functions from the fig.4.2. This test

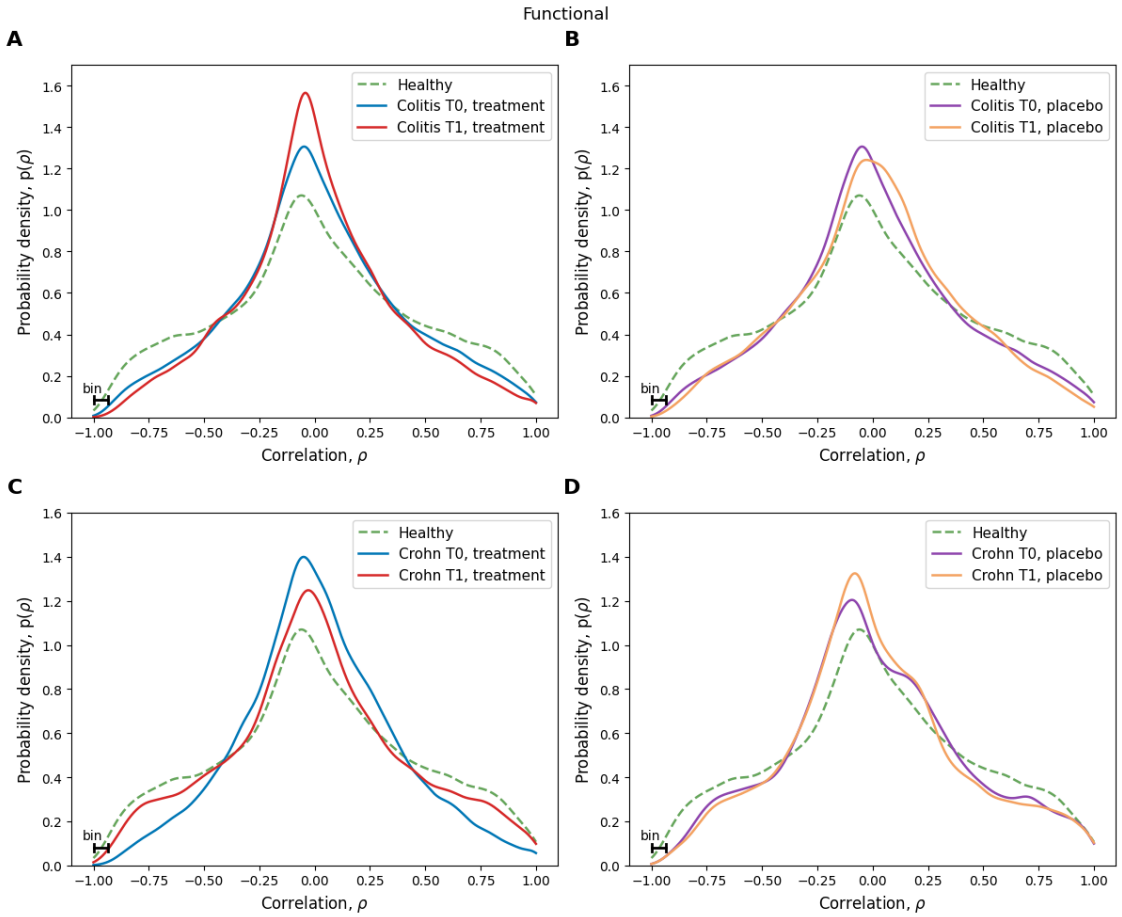

Figure 4.2: Histograms of the values of the function/function correlation matrix before (T_0) and after (T_1) treatment. HD group put as a reference. (**A**, **B**) UC patients, treatment/placebo. (**C**, **D**) CD patients, treatment/placebo.

is a nonparametric test, which is used in order to determine whether two samples come from the same distribution. Adopted hypotheses:

- H_0: samples are drawn from the same distribution, $F(x) = G(x)$;
- H_1: samples are drawn from different distributions, $F(x) \neq G(x)$.

The KS statistic is defined as:

$$D_{n,m} = sup_x |F(x)_n - G(x)_m|,$$

where $F_n(x), G_m(x)$ are the empirical cumulative distribution functions (cdfs). For large

samples, H_0 is rejected at level α, if

$$D_{n,m} > c(\alpha)\sqrt{\frac{n+m}{n \times m}}. \tag{4.1}$$

The KS test was performed with the use of SciPy's function *ks_2samp* for two cases: first, the cdfs "before" and "after" treatment were compared; second, all the cdfs for IBD patients were compared to the cdf corresponding to HD. The results are shown in tab.4.1 and tab.4.2.

Table 4.1: Kolmogorov-Smirnov test for samples before and after treatment.

Samples	**KS-statistic**	**p-value**	**Conclusion**
UC: treatment	0.03	0.76	$F_{\text{before}} = G_{\text{after}}$
UC: placebo	0.021	0.98	$F_{\text{before}} = G_{\text{after}}$
CD: treatment	0.071	**0.013**	$F_{\text{before}} = G_{\text{after}}$
CD: placebo	0.014	0.99	$F_{\text{before}} = G_{\text{after}}$

Table 4.2: Kolmogorov-Smirnov test for samples for IBD patients and healthy donors.

Samples	**KS-statistic**	**p-value**	**Conclusion**
UC: treatment, T_0	0.063	0.04	$F_{\text{UC\_tr\_T}_0} \neq G_{\text{HD}}$
UC: treatment, T_1	0.083	0.002	$F_{\text{UC\_tr\_T}_1} \neq G_{\text{HD}}$
UC: placebo, T_0	0.063	0.04	$F_{\text{UC\_pl\_T}_0} \neq G_{\text{HD}}$
UC: placebo, T_1	0.066	0.03	$F_{\text{UC\_pl\_T}_1} \neq G_{\text{HD}}$
CD: treatment, T_0	0.097	**0.00016**	$F_{\text{CD\_tr\_T}_0} \neq G_{\text{HD}}$
CD: treatment, T_1	0.033	**0.65**	$F_{\text{CD\_tr\_T}_1} = G_{\text{HD}}$
CD: placebo, T_0	0.048	**0.2**	$F_{\text{CD\_pl\_T}_0} = G_{\text{HD}}$
CD: placebo, T_1	0.058	**0.07**	$F_{\text{CD\_pl\_T}_1} = G_{\text{HD}}$

At level $\alpha = 0.05$, as can be seen from tab.4.1, the null hypothesis cannot be rejected for all the cases. However, only for CD patients with sodium butyrate treatment the p-value is not far from the threshold and significantly lower than for other cases. Therefore, the conclusion that for CD patients, who received treatment, the histograms of correlation values changed much greater than in other cases can be drawn. This conclusion is in agreement with the previously written observations.

Next, regarding the results shown in tab.4.2, a stronger conclusion can be drawn: the sodium butyrate treatment made function/function correlation values for CD patients distributed the same as for HD. Therefore, the sodium butyrate treatment can be considered effective for CD patients. However, this table also brings another observation: "before" and "after" treatment placebo subset of CD patients' correlation values are drawn from the same distribution as for HD. In fact, this can be observed from fig.4.2(**C,D**), as the placebo subset from the beginning was much more resembling the healthy one than the treatment one. But, with the combination of the previous results, since there is no significant difference between two histograms for placebo subset, the conclusion of the positive influence of sodium butyrate treatment on CD patients remains.

Secondly, the decay profile of the eigenvalues of the correlation matrices was explored. According to [46], this profile should be steeper for healthy microbiome. This pattern is observed for both UC and CD (fig.4.3). Moreover, for CD patients who received treatment, the profile of eigenvalue decay became more resembling the one from HD, meaning that this is yet another signature of sodium butyrate treatment working.

The eigenvalues of the correlation matrices were studied further: the raised question was whether the spectrum of eigenvalues of functional correlation matrices differs from the one obtained from random matrices. The theoretical distribution for the eigenvalues of the random matrix is given by the Marchenko-Pastur theorem:

Theorem 4.1.1 *Let X be an $n \times m$ random matrix with independent, identically distributed (i.i.d.) entries with mean 0 and variance σ^2. Define the sample covariance matrix:*

$C = \frac{1}{n} XX^T$, which has dimension $n \times n$.

The eigenvalues of C follow the Marchenko-Pastur density (as $n, m \to \infty$ with a fixed ratio $q = \frac{m}{n}$):

$\rho(\lambda) = \frac{1}{2\pi q \sigma^2 \lambda} \sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}, \quad \lambda \in [\lambda_-, \lambda_+]$,

where:

- $\lambda_+ = \sigma^2(1 + \sqrt{q})^2$ is the upper limit of the spectrum,

- $\lambda_- = \sigma^2(1 - \sqrt{q})^2$ is the lower limit of the spectrum,

- $q = \frac{m}{n}$ is the aspect ratio of the matrix.

The properties are:

1. Spectrum Bounds: Most eigenvalues lie in the interval $[\lambda_-, \lambda_+]$. Eigenvalues outside this interval are considered outliers or noise.

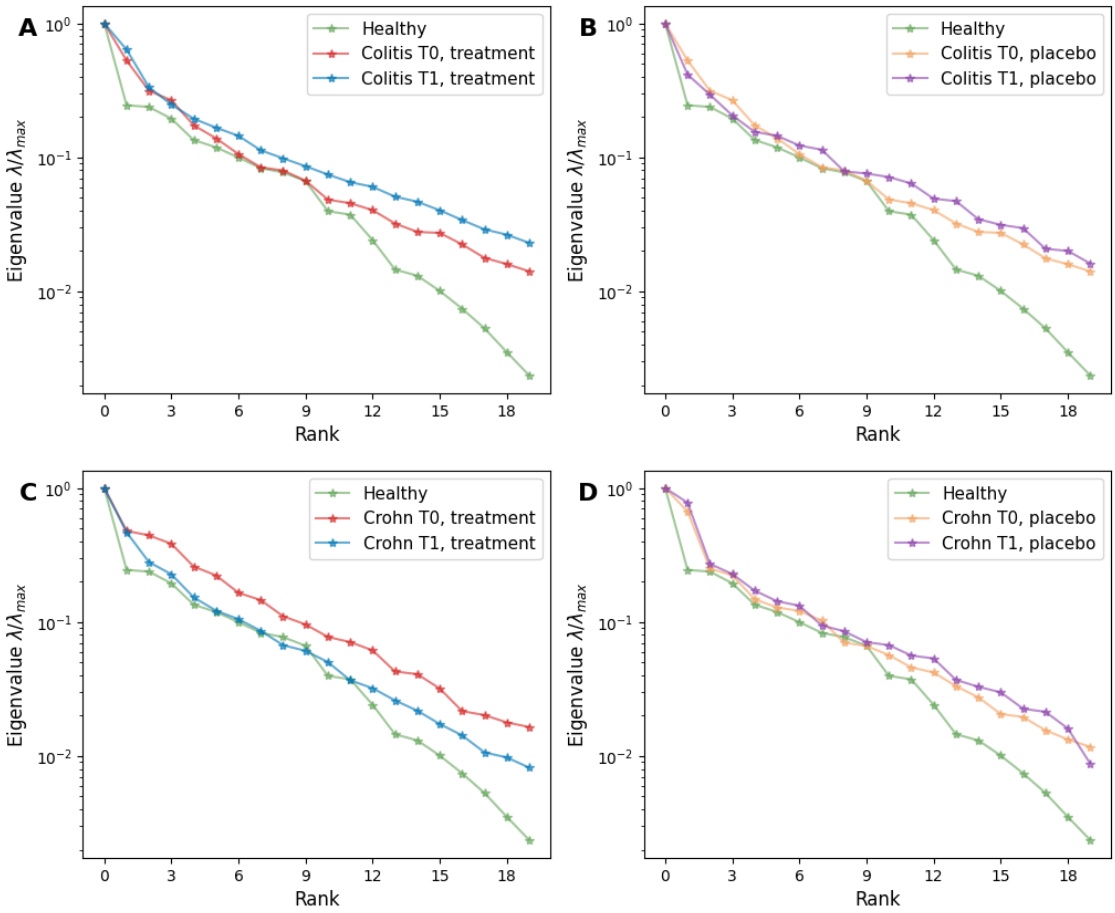

Figure 4.3: Decay profile of the eigenvalues before (T_0) and after (T_1) treatment. HD group put as a reference. (**A**, **B**) UC patients, treatment/placebo. (**C**, **D**) CD patients, treatment/placebo.

2. Limiting Cases: When $q \to 0$ (i.e., $m \gg n$), the distribution approaches a Dirac delta at σ^2.

As the relationship between correlation and covariance is given by $\rho_{xy} = \dfrac{cov(x,y)}{\sigma_x \sigma_y}$, it is possible to proceed with eigenvalues for the correlation matrices with applied normalization. The results are shown in fig.4.4 and fig.4.4. In each case, only eigenvalues in the range [0,12] or [0,10] are taken into account, however, for each correlation matrix there is a small subset of much larger eigenvalues (up to $\approx$ 120). From the figures it can be seen that for all the cases the majority of the eigenvalues are small and do not follow the Marchenko-Pastur distribution. First, this means that the inner structure of correlation matrices is far from random. Second, the fact that the majority of eigenvalues are small

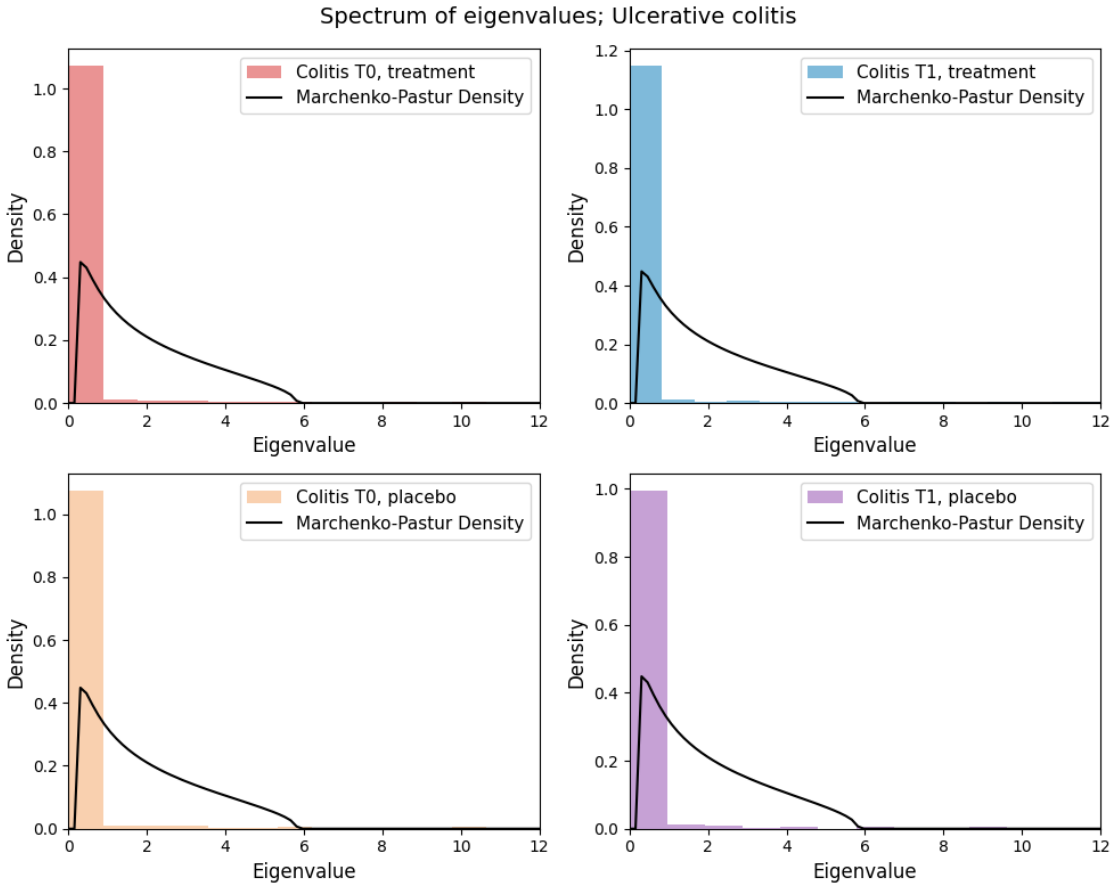

Figure 4.4: Spectrum of eigenvalues of function/function correlation matrices compared to the theoretical distribution for eigenvalues of random matrices: UC patients subset.

means that functions considered are redundant and/or the data is noisy. In the context of the adopted functional annotations, KEGG pathways, the redundancy can mean that only a subset of the pathways are activated in the considered conditions.

Lastly, the mean abundance (MA) of each function across UC and CD patients was found (with the division to treatment/placebo and two time points, as before). The result is shown in fig.4.6. As it can be observed from this figure, there is no distinct pattern for UC patients, as both treatment and placebo subsets have increase/decrease in functional MA. On the contrary, for CD patients the MAs for placebo subset stay the same before and after treatment, while for the group that received treatment some of the functional MAs increase and some decrease. In fig.4.6(**C**) two main groups of functions are separated for the treatment subset of CD patients for biological interpretation.

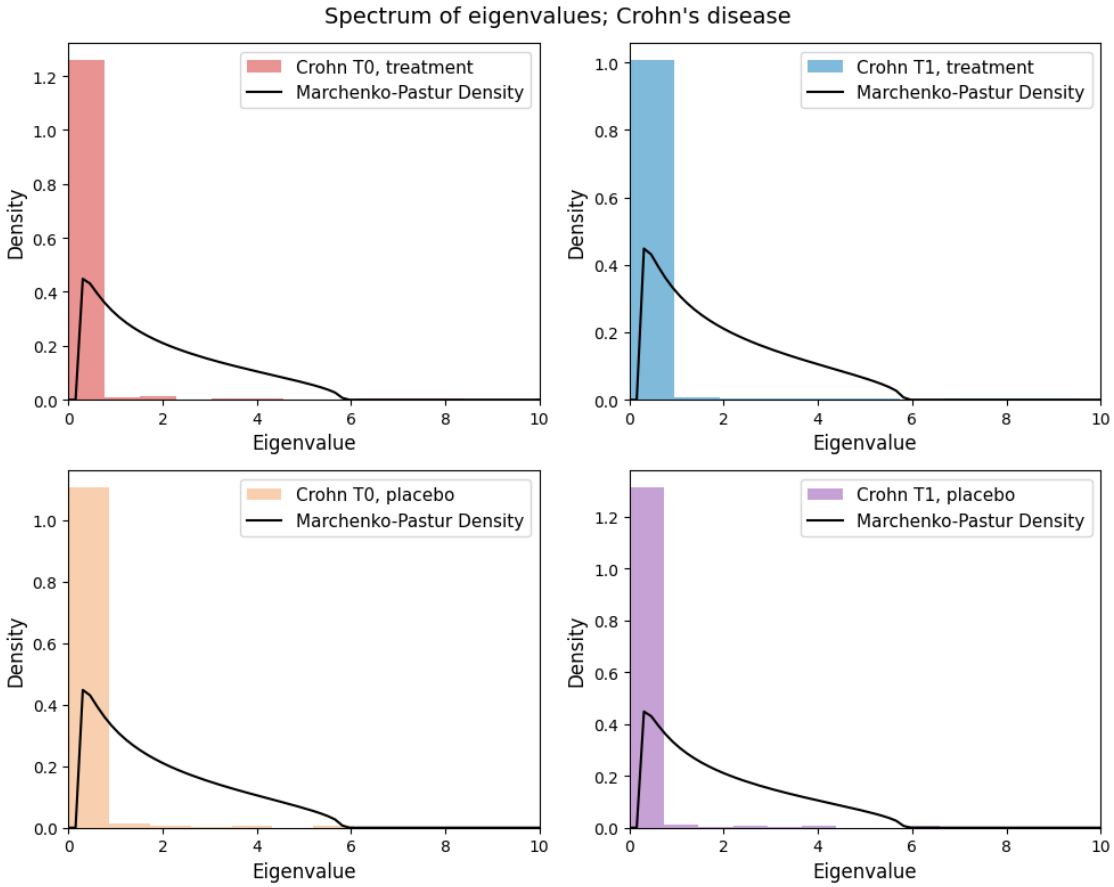

Figure 4.5: Spectrum of eigenvalues of function/function correlation matrices compared to the theoretical distribution for eigenvalues of random matrices: CD patients subset.

4.2 Biological interpretation of change in functional mean abundances for CD patients

In this section the two subsets of functions from 4.6(**C**) are discussed and interpreted with the regard of the previous studies on the butyrate [52].

As was mentioned above, KEGG-pathway was adopted as a functional annotation in this work. With the use of `https://www.kegg.jp/kegg/pathway.html`, it was possible to restore the names and classes of increased and decreased functions. The summary of the three most represented classes is provided in tab.4.3.

First, the three main classes of the decreased functions were explored. Regarding the most represented class, signal transduction, there is an established knowledge that bu-

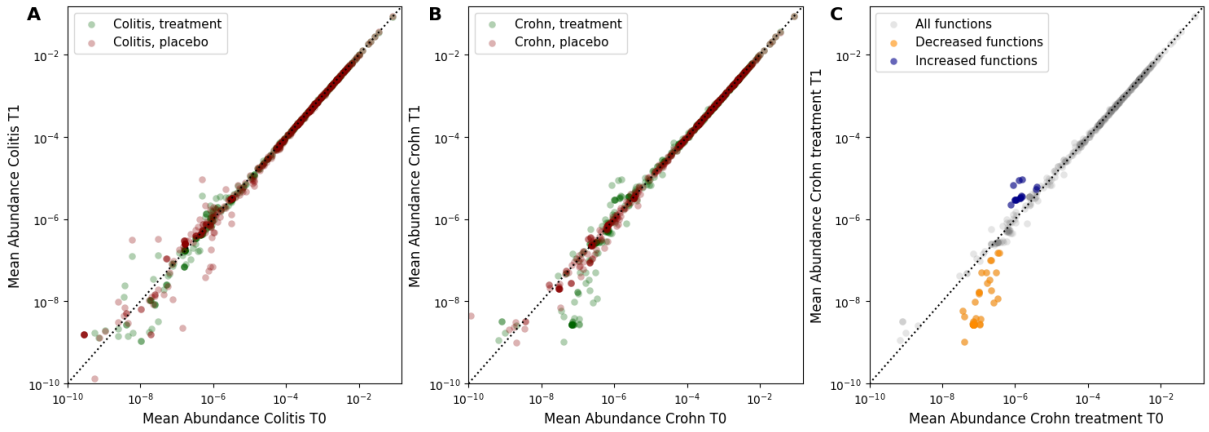

Figure 4.6: (**A**, **B**) Mean abundance of each function computed for UC and CD. Here the mean abundance before (T_0) and after (T_1) treatment are plotted on x and y axis respectively. The dotted line shows when before and after treatment subgroups have the same abundances. Functions above the dotted line increased after treatment, while the ones below decreased. (**C**) Increased and decreased functions for CD treatment subgroup.

Table 4.3: Three most represented classes of functions with changed mean abundances for CD patients who received sodium butyrate treatment.

Top	Decreased functions	Increased functions
1	Environmental Information Processing: Signal transduction	Human Diseases; Infectious disease:viral
2	Organismal Systems: Endocrine system	Cellular Processes: Cell growth and death
3	Organismal Systems: Nervous system	Human Diseases: Cancer:specific types

tyrate interacts with several signaling pathways associated with the maintenance of the intestinal barrier's integrity; specifically, with PI3K/Akt/mTOR [53, 54] and WNT/ERK [55] pathways. The results suggest that in CD, butyrate may modulate these signaling pathways as well. As for the endocrine system class, butyrate has been linked to the regulation of lipolysis in adipocytes [56] and ovarian steroidogenesis [57]. The results indicate that butyrate may participate in regulating these processes in CD, as we confirm potential involvement of these metabolic pathways in this condition. Regarding nervous system, butyrate has been found to influence the cholinergic and nitrergic phenotype of enteric neurons in the colon of rats [58], as well as it has been linked to the modulation of neurons such as orexin, which are associated with the endocannabinoid system [59]. The

results of this work suggest that butyrate may influence the activity of the enteric nervous system in CD, potentially affecting neural signaling.

Second, increased functions were investigated. As for the viral infectious disease class, butyrate-producing bacteria was associated with a reduced risk of hospitalization for infections; hence the findings of this work highlight a relationship between butyrate and the regulation of signaling pathways associated with viral infections in patients with CD, suggesting a potential protective role. Regarding cancer type class, there was established that patients with IBD have an increased risk of developing colorectal cancer [60]. The results demonstrate a possible association between butyrate treatment and functions related to colorectal cancer, indicating that butyrate could play a role in modulating pathways linked to cancer progression in CD patients. Finally, as for the last most represented class, cell growth and death, the findings of this work support involvement of butyrate in modulating these cellular processes, as reported in [61, 62].

5 Conclusion

In this thesis, the functional patterns of the gut bacteria in IBD patients was explored using data-driven approaches. Besides that, the influence of sodium butyrate treatment on IBD patients was addressed.

First of all, the functional patterns of healthy/unhealthy gut microbiome, reported in a previous study ([46]), was found for the chosen dataset as well. For both Crohn's disease and ulcerative colitis patients the histograms of function/function correlation values are much lower near +1 and -1 than found in a healthy microbiome. Furthermore, decay profile of the correlation matrices eigenvalues follows the expected differences between healthy and unhealthy microbiomes. Thus, the results of this thesis align with the previous understanding that in a healthy microbiome there is a core of strongly correlated functions, while in IBD there is none.

Second, the approach discussed in this thesis provides a demonstrative way of evaluating treatment, alternating gut microbiome and thus its functional composition. As an addition to the statistical tests used in order to study treatment's efficiency, this approach could be used. Using this approach, it was found that the sodium butyrate treatment is effective for Crohn's disease patients, while it does not show any improvements for ulcerative colitis patients.

Third, the spectrum of the eigenvalues of the function/function correlation matrices was found to be non-identical to the spectrum of random matrices, thus the inner structures of these matrices are different. It was shown that possibly the considered functions are

highly redundant in the gut microbiome.

Lastly, since in this work the functional composition of the gut microbiome is discussed, it was possible to find the main classes of functions, that were changed after the sodium butyrate treatment for Crohn's disease patients. Using the results of functional analysis, it was found that butyrate treatment is possibly influencing processes linked to the signal transduction, endocrine and nervous system, infectious diseases and cancer, cellular processes in Crohn's disease.

Future work could be possibly focusing on the further exploration of functional patterns found in IBD, with the aim of finding of unique signatures of ulcerative colitis and Crohn's disease. Furthermore, more detailed studies can be focused on the influence of the sodium butyrate treatment on the Crohn's disease.

References

[1] Qingdong Guan. “A Comprehensive Review and Update on the Pathogenesis of Inflammatory Bowel Disease”. In: *Journal of Immunology Research* 2019.1 (2019), p. 7247238.

[2] Giulia Roda, Siew Chien Ng, Paulo Gustavo Kotze, et al. “Crohn’s disease”. In: *Nature Reviews Disease Primers* 6.1 (2020), p. 22.

[3] Siew C Ng, Hai Yun Shi, Nima Hamidi, et al. “Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies”. In: *The Lancet* 390.10114 (2017), pp. 2769–2778.

[4] Atsushi Nishida, Ryo Inoue, Osamu Inatomi, et al. “Gut microbiota in the pathogenesis of inflammatory bowel disease”. In: *Clinical journal of gastroenterology* 11 (2018), pp. 1–10.

[5] Elizabeth Thursby and Nathalie Juge. “Introduction to the human gut microbiota”. en. In: *Biochem. J.* 474.11 (May 2017), pp. 1823–1836.

[6] Celeste Allaband, Daniel McDonald, Yoshiki Vázquez-Baeza, et al. “Microbiome 101: Studying, analyzing, and interpreting gut microbiome data for clinicians”. en. In: *Clin. Gastroenterol. Hepatol.* 17.2 (Jan. 2019), pp. 218–230.

[7] Perrine Hugon, Jean-Charles Dufour, Philippe Colson, et al. “A comprehensive repertoire of prokaryotic species identified in human beings”. In: *The Lancet Infectious Diseases* 15.10 (Oct. 2015), pp. 1211–1219.

[8] Junhua Li, Huijue Jia, Xianghang Cai, et al. “An integrated catalog of reference genes in the human gut microbiome”. In: *Nature Biotechnology* 32.8 (Aug. 2014), pp. 834–841.

[9] Junhua Li, Huijue Jia, Xianghang Cai, et al. “An integrated catalog of reference genes in the human gut microbiome”. en. In: *Nat. Biotechnol.* 32.8 (Aug. 2014), pp. 834–841.

[10] Sylvia H Duncan, Elena Conti, Liviana Ricci, et al. “Links between diet, intestinal anaerobes, microbial metabolites and health”. en. In: *Biomedicines* 11.5 (May 2023).

[11] Baohong Wang, Mingfei Yao, Longxian Lv, et al. “The Human Microbiota in Health and Disease”. In: *Engineering* 3.1 (2017), pp. 71–82. ISSN: 2095-8099. DOI: https://doi.org/10.1016/J.ENG.2017.01.008. URL: https://www.sciencedirect.com/science/article/pii/S2095809917301492.

[12] Petra Louis, Georgina L Hold, and Harry J Flint. “The gut microbiota, bacterial metabolites and colorectal cancer”. In: *Nature reviews microbiology* 12.10 (2014), pp. 661–672.

[13] Jean Guy LeBlanc, Christian Milani, Graciela Savoy De Giori, et al. “Bacteria as vitamin suppliers to their host: a gut microbiota perspective”. In: *Current opinion in biotechnology* 24.2 (2013), pp. 160–168.

[14] Andreas J Bäumler and Vanessa Sperandio. “Interactions between the microbiota and pathogenic bacteria in the gut”. In: *Nature* 535.7610 (2016), pp. 85–93.

[15] Peng Qiu, Takatsugu Ishimoto, Lingfeng Fu, et al. “The Gut Microbiota in Inflammatory Bowel Disease”. In: *Frontiers in Cellular and Infection Microbiology* 12 (2022). ISSN: 2235-2988. DOI: 10.3389/fcimb.2022.733992. URL: https://www.frontiersin.org/journals/cellular-and-infection-microbiology/articles/10.3389/fcimb.2022.733992.

[16] Josephine Ni, Gary D Wu, Lindsey Albenberg, et al. “Gut microbiota and IBD: causation or correlation?” In: *Nature reviews Gastroenterology & hepatology* 14.10 (2017), pp. 573–584.

[17] Helal F. Hetta, Yasmin N. Ramadan, Ahmad A. Alharbi, et al. “Gut Microbiome as a Target of Intervention in Inflammatory Bowel Disease Pathogenesis and Therapy”. In: *Immuno* 4.4 (2024), pp. 400–425. ISSN: 2673-5601. URL: https://www.mdpi.com/2673-5601/4/4/26.

[18] Rebecca J Case, Yan Boucher, Ingela Dahllöf, et al. “Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies”. en. In: *Appl. Environ. Microbiol.* 73.1 (Jan. 2007), pp. 278–288.

[19] J. Michael Janda and Sharon L. Abbott. "16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls". In: *Journal of Clinical Microbiology* 45.9 (2007), pp. 2761–2764. DOI: 10.1128/jcm.01228-07. eprint: https://journals.asm.org/doi/pdf/10.1128/jcm.01228-07. URL: https://journals.asm.org/doi/abs/10.1128/jcm.01228-07.

[20] Richa Bharti and Dominik G Grimm. "Current challenges and best-practice protocols for microbiome analysis". In: *Briefings in Bioinformatics* 22.1 (Dec. 2019), pp. 178–193. ISSN: 1477-4054. DOI: 10.1093/bib/bbz155. eprint: https://academic.oup.com/bib/article-pdf/22/1/178/35934895/bbz155.pdf. URL: https://doi.org/10.1093/bib/bbz155.

[21] CD Genomics. *Introduce to 16S rRNA and 16S rRNA Sequencing*. 2024. URL: https://www.cd-genomics.com/blog/introduce-to-16s-rrna-and-16s-rrna-sequencing/ (visited on 11/11/2024).

[22] Loren A Honaas, Naomi S Altman, and Martin Krzywinski. "Study design for sequencing studies". In: *Statistical genomics: Methods and protocols* (2016), pp. 39–66.

[23] Marlène Chiarello, Mark McCauley, Sébastien Villéger, et al. "Ranking the biases: The choice of OTUs vs. ASVs in 16S rRNA amplicon data analysis has stronger effects on diversity measures than rarefaction and OTU identity threshold". In: *PloS one* 17.2 (2022), e0264443.

[24] Moira Marizzoni, Thomas Gurry, Stefania Provasi, et al. "Comparison of bioinformatics pipelines and operating systems for the analyses of 16S rRNA gene amplicon sequences in human fecal samples". en. In: *Front. Microbiol.* 11 (June 2020), p. 1262.

[25] Patrick D. Schloss, Sarah L. Westcott, Thomas Ryabin, et al. "Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities". In: *Applied and Environmental Microbiology* 75.23 (2009), pp. 7537–7541. ISSN: 0099-2240. DOI: 10.1128/AEM.01541-09. URL: https://aem.asm.org/content/75/23/7537.

[26] *Overview of QIIME 2 Plugin Workflows*. https://docs.qiime2.org/2024.10/tutorials/overview/. Accessed: 2024-11.

[27] Benjamin J Callahan, Paul J McMurdie, Michael J Rosen, et al. “DADA2: High-resolution sample inference from Illumina amplicon data”. en. In: *Nat. Methods* 13.7 (July 2016), pp. 581–583.

[28] Evan Bolyen, Jai Ram Rideout, Matthew R. Dillon, et al. “Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2”. In: *Nature Biotechnology* 37.8 (Aug. 2019), pp. 852–857. ISSN: 1546-1696. DOI: 10.1038/s41587-019-0209-9. URL: https://doi.org/10.1038/s41587-019-0209-9.

[29] Christian Quast, Elmar Pruesse, Pelin Yilmaz, et al. “The SILVA ribosomal RNA gene database project: improved data processing and web-based tools”. In: *Nucleic Acids Research* 41.D1 (Nov. 2012), pp. D590–D596. ISSN: 0305-1048. DOI: 10.1093/nar/gks1219. eprint: https://academic.oup.com/nar/article-pdf/41/D1/D590/3690367/gks1219.pdf. URL: https://doi.org/10.1093/nar/gks1219.

[30] T Z DeSantis, P Hugenholtz, N Larsen, et al. “Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB”. en. In: *Appl. Environ. Microbiol.* 72.7 (July 2006), pp. 5069–5072.

[31] Minoru Kanehisa and Susumu Goto. “KEGG: Kyoto Encyclopedia of Genes and Genomes”. In: *Nucleic Acids Research* 28.1 (Jan. 2000), pp. 27–30. ISSN: 0305-1048. DOI: 10.1093/nar/28.1.27. eprint: https://academic.oup.com/nar/article-pdf/28/1/27/9895154/280027.pdf. URL: https://doi.org/10.1093/nar/28.1.27.

[32] Jaime Huerta-Cepas, Damian Szklarczyk, Davide Heller, et al. “eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses”. en. In: *Nucleic Acids Res.* 47.D1 (Jan. 2019), pp. D309–D314.

[33] Michael Y Galperin, Yuri I Wolf, Kira S Makarova, et al. “COG database update: focus on microbial diversity, model organisms, and widespread pathogens”. en. In: *Nucleic Acids Res.* 49.D1 (Jan. 2021), pp. D274–D281.

[34] Elodie Drula, Marie-Line Garron, Suzan Dogan, et al. “The carbohydrate-active enzyme database: functions and literature”. en. In: *Nucleic Acids Res.* 50.D1 (Jan. 2022), pp. D571–D577.

[35] Ron Caspi, Tomer Altman, Richard Billington, et al. “The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases”. In: *Nucleic Acids Research* 42.D1 (Nov. 2013), pp. D459–D471. ISSN: 0305-1048. DOI: 10.1093/nar/gkt1103. eprint: https://academic.oup.com/nar/article-pdf/42/D1/D459/3578456/gkt1103.pdf. URL: https://doi.org/10.1093/nar/gkt1103.

[36] Robert D Finn, Alex Bateman, Jody Clements, et al. “Pfam: the protein families database”. en. In: *Nucleic Acids Res.* 42.Database issue (Jan. 2014), pp. D222–30.

[37] Christophe Djemiel, Pierre-Alain Maron, Sébastien Terrat, et al. “Inferring microbiota functions from taxonomic genes: a review”. In: *GigaScience* 11 (Jan. 2022), giab090. ISSN: 2047-217X. DOI: 10.1093/gigascience/giab090. eprint: https://academic.oup.com/gigascience/article-pdf/doi/10.1093/gigascience/giab090/42276972/giab090_reviewer_2_report_original_submission.pdf. URL: https://doi.org/10.1093/gigascience/giab090.

[38] Gavin M Douglas, Vincent J Maffei, Jesse R Zaneveld, et al. “PICRUSt2 for prediction of metagenome functions”. In: *Nature biotechnology* 38.6 (2020), pp. 685–688.

[39] Franziska Wemheuer, Jessica A Taylor, Rolf Daniel, et al. “Tax4Fun2: prediction of habitat-specific functional profiles and functional redundancy based on 16S rRNA gene sequences”. In: *Environmental Microbiome* 15 (2020), pp. 1–12.

[40] Sonia Facchin, Nicola Vitulo, Matteo Calgaro, et al. “Microbiota changes induced by microencapsulated sodium butyrate in patients with inflammatory bowel disease”. In: *Neurogastroenterology & Motility* 32.10 (2020), e13914.

[41] Sonia Facchin, Matteo Calgaro, Mattia Pandolfo, et al. *Oral Butyrate Improves Clinical Outcome in IBD: A Randomised Placebo-Controlled Study Targeting Gut Microbiota*. Submitted for publication to American Journal of Gastroenterology.

[42] Marcel Martin. “Cutadapt removes adapter sequences from high-throughput sequencing reads”. In: *EMBnet. journal* 17.1 (2011), pp. 10–12.

[43] Michael S Robeson, Devon R O’Rourke, Benjamin D Kaehler, et al. “RESCRIPt: Reproducible sequence taxonomy reference database management”. In: *PLoS computational biology* 17.11 (2021), e1009581.

[44] Nicholas A Bokulich, Benjamin D Kaehler, Jai Ram Rideout, et al. "Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin". In: *Microbiome* 6 (2018), pp. 1–17.

[45] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, et al. "Scikit-learn: Machine learning in Python". In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830.

[46] Marcello Seppi, Jacopo Pasqualini, Sonia Facchin, et al. "Emergent functional organization of gut microbiomes in health and diseases". In: *Biomolecules* 14.1 (2023), p. 5.

[47] Alexandre Almeida, Stephen Nayfach, Miguel Boland, et al. "A unified catalog of 204,938 reference genomes from the human gut microbiome". In: *Nature biotechnology* 39.1 (2021), pp. 105–114.

[48] Jai Ram Rideout, Greg Caporaso, Evan Bolyen, et al. *scikit-bio/scikit-bio: scikit-bio 0.6.2*. 2024.

[49] Liang Tian, Xu-Wen Wang, Ang-Kun Wu, et al. "Deciphering functional redundancy in the human microbiome". In: *Nature communications* 11.1 (2020), p. 6217.

[50] Nicholas J. Higham. "Computing a nearest symmetric positive semidefinite matrix". In: *Linear Algebra and its Applications* 103 (1988), pp. 103–118. ISSN: 0024-3795. DOI: https://doi.org/10.1016/0024-3795(88)90223-6. URL: https://www.sciencedirect.com/science/article/pii/0024379588902236.

[51] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[52] Sonia Facchin. personal communication. Oct. 2024.

[53] Gang Tang, Yi Du, Haochen Guan, et al. "Butyrate ameliorates skeletal muscle atrophy in diabetic nephropathy by enhancing gut barrier function and FFA2-mediated PI3K/Akt/mTOR signals". In: *British Journal of Pharmacology* 179.1 (2022), pp. 159–178.

[54] GuoYan Wang, SenLin Qin, Lei Chen, et al. "Butyrate dictates ferroptosis sensitivity through FFAR2-mTOR signaling". In: *Cell death & disease* 14.4 (2023), p. 292.

[55] Liping Liang, Le Liu, Wanyan Zhou, et al. “Gut microbiota-derived butyrate regulates gut mucus barrier repair by activating the macrophage/WNT/ERK signaling pathway”. In: *Clinical Science* 136.4 (2022), pp. 291–307.

[56] Hideo Ohira, Yoshio Fujioka, Chikae Katagiri, et al. “Butyrate attenuates inflammation and lipolysis generated by the interaction of adipocytes and macrophages”. In: *Journal of atherosclerosis and thrombosis* 20.5 (2013), pp. 425–442.

[57] Qianhong Ye, Xiangfang Zeng, Shuai Wang, et al. “Butyrate drives the acetylation of histone H3K9 to activate steroidogenesis through PPARγ and PGC1α pathways in ovarian granulosa cells”. In: *The FASEB Journal* 35.2 (2021), e21316.

[58] Etienne Suply, Philine de Vries, Rodolphe Soret, et al. “Butyrate enemas enhance both cholinergic and nitrergic phenotype of myenteric neurons and neuromuscular transmission in newborn rat colon”. In: *American Journal of Physiology-Gastrointestinal and Liver Physiology* 302.12 (2012), G1373–G1380.

[59] Nicola Forte, Brenda Marfella, Alessandro Nicois, et al. “The short-chain fatty acid acetate modulates orexin/hypocretin neurons: A novel mechanism in gut-brain axis regulation of energy homeostasis and feeding”. In: *Biochemical Pharmacology* (2024), p. 116383.

[60] Ryan W Stidham and Peter DR Higgins. “Colorectal cancer in inflammatory bowel disease”. In: *Clinics in colon and rectal surgery* 31.03 (2018), pp. 168–178.

[61] Masumeh Sanaei, Fraidoon Kavoosi, and Mohsen Safari. “Effect of 5’-fluoro-2’-deoxycytidine and sodium butyrate on the gene expression of the intrinsic apoptotic pathway, p21, p27, and p53 genes expression, cell viability, and apoptosis in human hepatocellular carcinoma cell lines”. In: *Advanced Biomedical Research* 12.1 (2023), p. 24.

[62] Kishor Pant, Amit K Mishra, Saman Man Pradhan, et al. “Butyrate inhibits HBV replication and HBV-induced hepatoma cell proliferation via modulating SIRT-1/Ac-p53 regulatory axis”. In: *Molecular Carcinogenesis* 58.4 (2019), pp. 524–532.