Summer 6-15-2023

Metaproteomics Profiling of Gut Microbiota in Pediatric Acute Lymphoblastic Leukemia Patients

Shahd Ezz El Din
shahdezzeldin@aucegypt.edu

Graduate Studies

Metaproteomics Profiling of Gut Microbiota in Pediatric Acute Lymphoblastic Leukemia Patients

A THESIS SUBMITTED BY

Shahd Ezz El Din

TO THE

Biotechnology Graduate Program

SUPERVISED BY

Dr. Ahmed Moustafa

Biology Department
The American University in Cairo
Cairo, Egypt

Dr. Sameh Magdeldin

Proteomics and Metabolomics Research Program
Children Cancer Hospital 57357
Cairo, Egypt

In partial fulfillment of the requirements for the degree of
Master of Science in Biotechnology
2022

Declaration of Authorship

I, Shahd Ezz El Din, declare that this thesis titled, "**Metaproteomics Profiling of Gut Microbiota in Pediatric Acute Lymphoblastic Leukemia Patients**" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Shahd Ezz El Din

Date:

February 3, 2023

Approved by

Dr. Ahmed Mostafa (Supervisor) _____
Biology Department
The American University in Cairo

Dr. Sameh Magdeldin (Supervisor) _____
Proteomics and Metabolomics Research Program
Children Cancer Hospital 57357

Dr. Ahmed Abdellatif (Internal Examiner) _____
Biology Department
The American University in Cairo

Dr. Ramy Aziz (External Examiner) _____
Department of Microbiology and Immunology, Faculty of Pharmacy, Cairo University

Dr. Andreas Kakarougkas (Moderator) _____
Biology Department
The American University in Cairo

Abstract

The gut microbiota is the collection of microorganisms that harbor the human gastrointestinal tract. The relation of microbiota and cancer is an emerging field as several studies have linked the gut microbiota to cancer. Such findings opened the door towards a new era of studies aiming to reveal the role of the microbiota in cancer initiation and progression. Although several metagenomic studies have provided insights into the gut microbial composition in health and diseased state, the microbial functional characteristics is still poorly understood. With current advances in mass spectrometry, comprehensive understanding of microbiota proteome became possible. In this vein, the objective of this study was to explore the gut microbial composition in stool samples from pediatric acute lymphoblastic leukemia (ALL) patients, compared to the non-cancerous profile. The ALL is one of the most common pediatric cancer worldwide accounting for 25% of all childhood cancer cases, and identify the microbial proteins, and their functional implications in pediatric ALL patients. The microbiota profiling analysis revealed the microbial composition in the stool of ALL patients, including the high abundance of some genera such as *Bacteroides*, *Prevotella*, and *Streptococcus*, compared to other less abundant organisms, such as *Blautia, Lachnospiraceae, Roseburia*. Additionally, the metaproteomics results inferred the functional implications of the gut microbiota in ALL. The results pinpointed higher iron demand and oxidative stress in the stool of ALL patients compared to healthy individuals. Furthermore, functions related to amino acid, carbohydrate and butyrate metabolism were downregulated in ALL. These promising results are preliminary step toward a deeper understanding of the gut microbiota in ALL.

Acknowledgements

I would like to express my sincere gratitude to all my professors at The American University in Cairo for all the unconditional support throughout my master's journey. I would like to extend my gratitude to my co-advisor Dr. Sameh Magdeldin for his outstanding support. Also, I would like to thank my family and my dear friends.

Contents

Declaration of Authorship……..……....i

Abstract…….……...……iii

Acknowledgements…….....iv

List of Figures/Images/Sound Files……..……vi

List of Abbreviations……..…….....vii

Chapter 1. Introduction……..…….…….....1

1.1 Gut Microbiota……...……...…1

1.2 The Microbiota through Omics…….……2

1.3 Acute Lymphoblastic Leukemia…….……..4

1.4 Study Objective…….....……..5

Chapter 2. Materials and Method…….…….……6

2.1 Patients and Clinical Information……...……..6

2.2 DNA Extraction and Quantification……...…….…….6

2.3 16sRNA Sequencing….…….……...……7

2.4 Quality Control…….……...…8

2.5 16S rRNASequence Analysis……..…….……8

2.6 Phylogenetic Analysis of 16S rRNA……..……..…..8

2.7 Pseudo-metagenome Database……...……..…….8

2.8 Sample Processing and Protein Extraction…….…….8

2.9 Protein Precipitation and Quantification…….……9

2.10 In-gel Trypsin Digestion…….…..…….…….9

2.11 Nano LC-MS/MS Analysis…….…….……10

2.12 Metaproteomics Data Analysis…….…….……..10

2.13 Metaproteomics Statistical Analysis…….…….……...…11

2.14 Gene Ontology-based Functional Analysis……...…11

Chapter 3. Results……..12

3.1 Microbiota Profile of Patients with ALL…….……...….12

3.2 Microbial Proteins Revealed by Metaproteomics……..…….……..….19

3.3 Functional Assignment of Microbial Proteins in Patients with ALL.……..……..…23

Chapter 4. Discussion……...…….....28

Chapter 5. Conclusion……..…….....33

References……...34

Appendix 1…….....36

List of Figures/Images/Sound Files

Figure 1: Microbiota composition of the ALL samples on the domain level. The percentage of each domain is shown. 13
Figure 2: Microbiota composition of the ALL samples on the phylum level. The percentage of each phylum is shown 14
Figure 3: The relative abundance of the identified microbial composition in each ALL sample. The x-axis represents the samples, and the y-axis shows the relative frequency in percentage. The color legend of the phyla is shown. 15
Figure 4: The most prevalent classes of the Firmicutes phylum in the ALL samples. The percentage of each class is shown, calculated from the total Firmicutes. 16
Figure 5: The relative abundance of the identified microbial composition, at the class level, in each ALL sample. The x-axis represents the samples, and the y-axis shows the relative frequency in percentage. The color legend of the phyla is shown. 17
Figure 6: Principle component analysis of gut microbiota profile in the study cohorts. 17
Figure 7: Microbiota Composition at phyla level, in controls and ALL patients. The x-axis represents the median relative abundance. 18
Figure 8: Differentially abundant genera between the controls and ALL patients. 19
Figure 9: 2D principal component analysis (PCA) plot of the study cohorts, ALL and controls. PC1 and PC2 are the first and the second principal component, respectively. 20
Figure 10: Hierarchical clustering of study samples showing the expression pattern of the top 100 significant proteins between ALL, and control. Significant proteins are shown on the y-axis, study cohorts are shown on the x-axis, displaying clusters of samples to study groups. Relative expression of each protein shown based on the z-score of the protein's normalized NSAF, with overexpressed and downregulated proteins depicted in red and blue, respectively. 21
Figure 11: Volcano plot representing expression profile of proteins that differ significantly between pediatric ALL patients and control individuals. Horizontal grey line represents $-\log_{10}$(adjusted pvalue) of 1.3. Vertical grey lines represent a magnitude of either ± 1 $\log_2$ fold change. Red and blue dots represent significantly differentially expressed proteins (p-value < 0.05, FDR-adjusted p-value < 0.05 and $\log_2$ fold change $\geq \pm 1$). Non-significant proteins are shown as grey dots. 22
Figure 12: Gene Ontology (GO) annotations of the significant upregulated proteins in ALL compared to control. GO annotations, according to the Biological Process, Molecular Function, and Cellular Localization, are shown on the x-axis, whereas the y-axis represents the number of proteins. 24
Figure 13: TreeMap visualization of REVIGO showing the GO Biological Process annotations of the upregulated proteins in ALL. Rectangles with the same color represents one cluster. 25
Figure 14: Gene Ontology (GO) annotations of the significant downregulated proteins in ALL compared to control. GO annotations, according to the Biological Process, Molecular Function, and Cellular Localization, are shown on the x-axis, whereas the y-axis represents the number of proteins. 26
Figure 15: TreeMap visualization of REVIGO showing the GO Biological Process annotations of the downregulated proteins in ALL patients. Rectangles with the same color represents one cluster. 27
Figure 16. Gut microbiota in pediatric ALL patients 32

List of Abbreviations

Acetonitrile (ACN)
Acute Lymphoblastic Leukemia (ALL)
Base pairs (bp)
Differentially expressed proteins (DEPs)
Dithiothreitol (DTT)
False discovery rate (FDR)
Gene Ontology (GO)
Human gastrointestinal tract (GIT)
Human Intestinal Tract (Meta-HiT)
Human microbiome project (HMP)
Iodoacetamide (IAA)
Normalized spectral abundance factor (NSAF)
Principle Component Analysis (PCA)
Probabilistic quotient normalization (PQN)
Short-chain fatty acids (SCFAs)
T helper 17 cells (TH17)

Chapter 1

Introduction

1.1 Gut Microbiota

The human gastrointestinal tract (GIT) is inhabited by a huge diverse and complex population of microorganisms. These hundred trillion microorganisms are known as gut microbiota [1]. It is estimated that the human GIT is harbored by approximately 10^{14} microorganisms. Bacteria represents the major sector of the microbiota, other microorganisms include archaea, yeast and protozoa. The most abundant bacterial phyla are *Firmicutes*, *Bacteroidetes*, *Proteobacteria*, *Actinobacteria*, and *Fusobacteria*. *Firmicutes* and *Bacteroidetes* predominate, accounting for about 90% of the gut microorganisms. The *Firmicutes* phylum is mostly represented by the *Clostridium* genus, while *Bacteroides* and *Prevotella* are the abundant genera of *Bacteroidetes* [2]. At early stage, the microbial composition is affected by the mode of delivery and the method of milk feeding, and is highly altered until the age of three years when it becomes relatively stable resembling that of the adults [2]. Furthermore, the composition of gut microbiota is highly variable from one person to another depending on several factors, such as age, mood, and diet [3]. The latter is an important factor leading to differences in microbial composition from one individual to another, or from one community to another. For example, the *Prevotella* genus was found abundant in African children in individuals with high carbohydrates uptake [3]. Additionally, the microbiota was found to be individual specific [4], yet, some bacterial taxa were found to be shared across individuals [1]. These species are referred to as core microbiota and represents about 10% of the human microbiota [5].

Due to its vast majority and diversity, the microbiota plays a crucial role in the human health, including vitamins synthesis [3]. For instance, *Bifidobacteria* genus, from the *Actinomycetota* phylum, produces folate vitamin required for the formation of red blood cells [3]. Additionally, gut microbiota could

contribute to the synthesis of essential vitamins that are not secreted by the human body, e.g. *de novo* vitamins, such as vitamin B12 produced mainly by lactic acid bacteria. Furthermore, gut microbiota contributes to the maintenance of the gut mucosal barrier integrity by being the dominant source of short-chain fatty acids (SCFAs) production [3]. SCFAs are fatty acids synthesized via the fermentation of undigested polysaccharides by the microbiota [6].In the human gut, the most dominant SCFAs are acetate, butyrate, and propionate, in a ratio 3:1:1, respectively [3]. Butyrate, produced mainly by *Firmicutes*, is essential for the integrity of gut epithelial, in addition to its role as an anti-inflammatory. Propionate, secreted by *Bacteroidetes*, is involved in gluconeogenesis initiation, and hence, is required for glucose homeostasis [3]. Moreover, the SCFA acetate has a protective function through the blockage of the pathogenic *Escherichia coli* toxins [5]. Furthermore, gut microbiota interacts massively with the human host to maintain immune homeostasis. For instance, microbiota can protect the host from pathogenic species through several lines of immune defense including T-helper cells and immunoglobulin A [5]. Moreover, segmented filamentous bacteria secrete flagellins triggering T-helper cells through Toll-like receptor 5, mediating immune responses, both innate and adaptive. Furthermore, IgA production could be induced by microbiota through Toll-like receptor signaling. The secreted IgA by its turn protects the gut epithelial from pathogenic bacteria [5]. The above mentioned SCFAs also play role in the host immunity through the production of interleukin-18 required for epithelial barrier homeostasis [3].

1.2 The Microbiota through Omics

Studying microbiota was limited as the majority could not be cultured [7]. With the advances in phylogenetic analysis, using 16S rRNA and whole genome sequencing techniques, characterization of microbiota within the human body was possible. Studying the genes of microbiota is known as metagenomics [8], and the term microbiome is used to refer to the whole genome of microbiota [1]. The microbiome harbors 100-fold genes more than the human genome, highlighting the significance of such microorganisms [9]. Several metagenomics studies have extensively attempted to characterize the microbiota and define the symbiotic relationship between human host and microbiota. One of the

remarkable metagenomics project was the human microbiome project (HMP) established in 2007 [10] and the European project “Metagenomics of the Human Intestinal Tract” (Meta-HiT) [11]. Both projects sequenced the human microbiota at different body sites including the gut, skin, and mouth. Combined data from these two projects generated a catalog of reference human gut microbial genes [12].

During the past years, studies were focused on the commensal relation between microbiota and human, however, recent evidences have observed alterations in the microbial composition, known as dysbiosis [3]. Several factors can cause microbial dysbiosis, including antibiotics, and specifically broad-spectrum antibiotics [2]. The latter disrupts the balance and alters the *Firmicutes* to *Bacteroidetes* ratio. Dysbiosis has linked to diseases including obesity [13], Crohn's disease [14], and type 2 diabetes [15]. Ferrer et al. showed elevation in *Firmicutes* and reduction in *Bacteroidetes* in obese individuals relative to lean subjects [13]. Crohn's patients were characterized by a significant decrease in *Faecalibacterium* genus compared to healthy individuals [14]. Surprisingly, microbial dysbiosis has been also linked to cancer [16]. In fact, microbial pathogens are thought to be the reason behind 15% to 20% of cancer cases [17]. Metagenomic analysis of cancer patients clarified that there are significant differences between gut microbial compositions compared to the healthy individuals. For instance, Dai et al. studied the gut microbiota in colorectal cancer patients from several countries through metagenomics analysis [18]. They found seven upregulated bacterial species associated with the cancer profile, such as *Bacteroides fragilis*, *Prevotella intermedia*, and *Fusobacterium nucleatum*. Additionally, another metagenomic study on breast cancer patients found differential microorganisms such as *Bifidobacterium cuniculi* and *Actinobacteria* [19]. Furthermore, the abundance of certain bacterial species such as *Fusobacterium nucleatum* was remarkably higher in oral squamous cell carcinoma patients than in healthy individuals [20]. Therefore, the relation between microbiota and cancer is a new emerging field, opening the door to further understand cancer pathophysiology and new cancer treatment strategies. Additionally, the crosstalk between symbiotic bacteria and host is still poorly understood.

Although metagenomic analysis has provided insights about gut microbial composition in health and

diseased state, there were several limitations. Analysis was limited to the identification of the existing microbial community and their gene sequence, with no clue on their precise function and their protein expression levels or their interaction with the host genome [21]. Moreover, metagenomics analysis is unable to differentiate between the active and dormant or dead microbiota [22]. Thus, scientists started to shift to other meta-omics approaches for further analysis of gut microbiota, among them metaproteomics, which is defined as the study of all protein samples recovered directly from environmental sources. In fact, metaproteomic analysis allows not only the identification of microorganisms but also their function [21]. It reveals information about the biological process, signal transduction, metabolic pathways, and protein expression. Moreover, metaproteomic analysis provides information about microbiota-host interactions [22]. With current advances in mass spectrometry, comprehensive understanding of the proteome of microbiota is now possible. There is a crucial need to characterize the role of the microbiota in cancer initiation, progression and treatment especially after a significant number of articles showing a tight link between microbiotal dysbiosis and the above mentioned diseases. For instance, gut microbiota in colorectal cancer has been also studied using metaproteomics analysis of fecal samples from colorectal cancer patients and healthy individuals [23]. Long et al. found 341 significantly different proteins related to oxidative stress, DNA replication, and iron transport.

1.3 Acute Lymphoblastic Leukemia

Recently, a metagenomics study characterizing the gut microbiota composition in Acute Lymphoblastic Leukemia (ALL) was done by Rajagopala et al. [24]. 16S rRNA analysis was done to fecal samples collected from 51 participants, divided into 23 patients and their corresponding healthy sibling, in addition to 5 patients without their healthy siblings. Some taxa were found in common between the control and the diseased cohorts, although the diseased profile showed significant lower microbial diversity than the control group. Taxa such as *Roseburia*, *Ruminococcus*, *Anaerostipes*, and *Coprococcus* showed decreased abundance in the diseased group. The abundance of several genera differs also between the control and the diseased group, such as *Bacteroides* and *Prevotella* with an abundance in the control group of 40.2% and

12.2% respectively, and 62.2% and 7.3% in the diseased group, respectively. Furthermore, in another study, next generation sequencing analysis of fecal samples showed significant differences in alpha and beta diversity between ALL patients and control group [25]. *Bacteroidetes* showed higher abundance in diseased group, confirming Rajagopala findings. In fact, ALL is one of the most common malignant cancer in children, accounting for more than 25% of all pediatric cancers, and 80% of leukemia's cases in children worldwide [26]. ALL is leukemia of the lymphocytes in the bone marrow, where lymphoid precursors start to proliferate uncontrollably replacing the normal hematopoietic cells of the bone marrow [24]. ALL is usually associated with chromosomal alterations and translocations, the most common is the t(9;22), also known as Philadelphia chromosome [27]. ALL is mainly classified into T lymphoblastic and B lymphoblastic, where the latter occupies 85% of the cases. The 5-years survival rate of ALL seems relatively high, yet the disease remains a leading death case in pediatric cancers due to the high relapse [27]. Additionally, patients receiving the combination chemotherapy treatment suffer from the treatment side effects.

1.4 Study Objective

In this vein, the objective of this study is to explore the diversity of gut microbiota in pediatric ALL patients and identify the microbial secreted proteins, and their impact on host-microbial interactions. Furthermore, insights on the differences between the healthy and the diseased profile could be achieved by comparing the patients' microbiota composition with the healthy group.

Chapter 2

Materials and Method

2.1 Patients and Clinical Information

Stool samples were obtained prospectively from newly diagnosed ALL pediatric patients at the Children's Cancer Hospital 57357 Egypt (CCHE-57357) following approval by the Institutional Research Board (IRB-CCHE-57357-35-2019), and the IRB committee at the American University in Cairo (AUC). Informed consent was obtained from patients' guardians. Thirty patients initially diagnosed with ALL, B-lineage were selected. Enrollment criteria were for patients above 3 years old and up to 18 years old, both genders, and patients with the 3 risk levels, low, standard and high. Patients did not receive any chemotherapy and/ or radiotherapy treatment, nor steroids at least one week prior to the study. Also, patients did not receive antibiotics two weeks prior to the study. Patients with other systemic diseases, such as cardiovascular, liver, respiratory and kidney diseases, and immune deficiency, and patients with trisomy 21 Down syndrome were excluded from the study. Regarding the control group, stool samples were collected from 30 healthy subjects following the same enrolling criteria as patients. Guardians of eligible controls were consented to join the study. The mean age of patients was 6.3 years (range 3-15 years), while the control was 7.4 years (range 3-13 years). The detailed sample descriptions are found in Supplementary Figure 1. All collected samples were stored at -80°C until use.

2.2 DNA Extraction and Quantification

DNA was extracted from 0.3 gram of each sample using PureLink™ Microbiome DNA Purification Kit (Catalog Number A29790, Invitrogen), following the manufacturer's instructions. Briefly, stool samples were mixed with lysis buffer and beads, and tubes were vortexed and incubated at 65 °C for 10 minutes. Samples were homogenized by bead beating at maximum speed for 10 minutes, and centrifuged at 14,000 × g for 5 minutes. A clean up step was done to eliminate inhibitors. Binding buffer was added to the

supernatant onto the spin column, which was centrifuged at 14,000 × g for 1 minute. One washing step was done followed by elution to collect the purified DNA. The DNA was quantified by DeNovix® dsDNA Assay, and the DNA quality was checked by NanoDrop instrument (Thermo Scientific).

2.3 16S rRNA Sequencing

Samples were sequenced by the Illumina MiSeq System. At first, amplification of the 16S rRNA gene was done using primers targeting the variable V3 and V4 regions [28]. The protocol creates a single amplicon of about 460 bp. A first PCR cycle was done, using 10 ng/µl microbial DNA, to amplify the V3 and V4 regions using the specific primers, at the following conditions; 95 °C for 3 minutes, followed by 25 cycles of 95 °C for 30 seconds, 55 °C for 30 seconds, and 72 °C for 30 seconds, followed by 72 °C for 5 minutes as a final extension step, then storage at 4 °C. The resulted fragment size was checked using the Bioanalyzer DNA 1000 chip with an expected size of about 550 bp. A PCR clean-up step was done using AMPure XP beads to purify the amplicon from unpaired primers. A second PCR was done to add the Illumina sequencing adapters and dual-index barcodes, i5 and i7, to the amplicon, using Nextera XT Index Kit, following similar run conditions as the first PCR, with the exception of 8 cycles rather than 25. A PCR clean-up step was done as well. The exact primers sequence, including the Illumina adapters, were as follows;

Forward Primer =

5'TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG

Reverse Primer =

5'GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC

The generated library was validated using Bioanalyzer DNA 1000 chip to verify the size, with an expected size of about 630 bp. Following PCR, libraries were quantified using dsDNA high sensitivity assay, normalized, and pooled, followed by sequencing on Illumina MiSeq.

2.4 Quality Control

The raw reads16S rRNA sequences, in fasta format, were retrieved and filtered based on reads length (≥300) and reads number (>0.2 Mb). Reads passing the filtration criteria were kept for downstream analysis.

2.5 16S rRNA Sequence Analysis

Filtered reads were processed by amplicon sequence variant (ASV) error correction with DADA2 platform, using QIIME2 [29]. Reads were processed to trim adapter sequences and chimeric sequences using the following parameters; trunc_len_f: 280, trunc_len_r: 240, trim_left_f: 0,and trim_left_r: 0. The ASV abundance tables were retrieved.

2.6 Phylogenetic Analysis of 16S rRNA

DNA sequences were mapped against reference database; Silva-based 16S (version 138), using q2-feature classifier plugin in QIIME2. A confidence cutoff of 0.7 was applied. A rooted phylogenetic tree was constructed. Each feature was assigned taxonomic rank, if available; phylum, class, order, family, genus, and species. The relative frequency of features in each sample was plotted as bar plot.

2.7 Pseudo-metagenome Database

Identified genera were used to create a pseudo database which was used as a reference database for the metaproteomics search. Protein sequences of these genera were downloaded from the Uniprot, both trEMBL and Swiss-Prot. Such workflow minimizes the time when searching against the whole Uniprot database, and limits as well the false discovery rate due to the huge search space [31].

2.8 Sample Processing and Protein Extraction

Stored stool samples at -80ºC will be thawed on ice. Microbial cells were separated via differential centrifugation to reduce the complexity of the samples. Briefly, 1.5-gram stool sample was suspended in 3 ml ice-cold PBS, and vigorously vortexed. Sample was homogenized for 15 minutes at tube rotator, and

centrifuged at 300xg, 4 °C for 5 minutes to collect supernatant. Supernatant was kept at 4°C. The pellet was subjected to two more rounds of resuspension in fresh PBS (2 ml/round) followed by low-speed centrifugation as described above. Finally, all collected supernatants were centrifuged at 14,000xg, 4 °C for 30 minutes. The resulted pellet was precipitated using 3-fold volume ice-cold acetone. Depending on pellet size, 2-3 ml lysis buffer (4% SDS in 100mM Tris-HCl pH 8.5, 1M TCEP, and 0.1% protease inhibitor cocktail) was be added. Sample was incubated at 95ºC for 15 minutes while shaking vigorously, followed by shaking on tube rotator for 30 minutes at room temperature. Sample was then centrifuged at 16,000xg for 15 min at 4°C. Pellet was discarded.

2.9 Protein Precipitation and Quantification

Supernatants were precipitated by 4-fold volume ice-cold acetone. The precipitated proteins were dissolved in 8M urea buffer (8M urea in 500mM Tris pH 8.5). Total protein concentration was determined by BCA assay using Pierce BCA Protein Assay Kit (Thermo Fisher Scientific, USA) according to the manufacturer's instructions [32].

2.10 In-gel Trypsin Digestion

Ninty µg of each quantified sample was subjected to in-gel digestion. In brief, samples were diluted with 2x Laemmli sample buffer supplemented with 5% β-mercaptoethanol, heated at 95°C for 10 minutes, and electrophoresed on 12.5% SDS-PAGE gel at 125 V for 90 minutes. Gels were stained with Coomassie Brilliant Blue G-250 Dye and destained using 12% methanol and 7% acetic acid solution, using the Invitrogen Power Blotter System (Thermo Fisher Scientific, USA). The gels were excised into ten pieces per lane and washed with 50 mM ammonium bicarbonate/50% acetonitrile (ACN) for 15 minutes/ 3 times. Gel pieces were then dried using a speed vacuum. Samples were reduced with 10 mM dithiothreitol (DTT) for 30 minutes at 60°C and alkylated with 55 mM Iodoacetamide (IAA) for 30 minutes at room temperature in the dark. Gel pieces were washed once with 25 mM ammonium bicarbonate, further cut more diminutive, and dehydrated with ACN for 15 minutes, followed by speed vacuuming. Samples were trypsinized with

10 ng/µL procaine trypsin (Sigma, Germany) at 37°C overnight (https://pubmed.ncbi.nlm.nih.gov/19381606/). Digested peptides were extracted with ACN: Milli-Q water: Formic acid in a ratio of 66:33:1, respectively, for 5 minutes twice. Speed vacuum was done, and peptides were reconstituted in 0.2% formic acid. Samples were sonicated for 10 minutes, and centrifuged at 10,000 rpm 8900 x g for 5 minutes prior to mass spectrometry injection.

2.11 Nano LC-MS/MS Analysis

Nano LC-MS/MS analysis was carried out using TripleTOF 5600+ mass spectrometer (Sciex, USA) coupled with Eksigent nanoLC-400 autosampler and Ekspert nanoLC 425 pump at the front end. A total volume of 10µl of the peptide solution was injected into the trap and elute mode. Peptides were loaded and trapped onto a 5 µm ChromXP C18-CL trap column, 10 x 0.5 mm (Sciex, USA). Peptides were then separated using a 3 µm ChromXP C18-CL reverse-phase column, 120Å, 150 x 0.3mm (Sciex, USA), at a 10 µl/min flow rate. Samples were subsequently eluted on a linear gradient 3-40% solution of 80% ACN and 0.2% formic acid for 55 minutes. In positive mode, the set ranges for MS and MS/MS were 400-1250 m/z and 170-1500 m/z, respectively. The top 40 intense peaks were sequentially selected in data-dependent acquisition (DDA) mode with the charge state 2-5. Full scan MS and MS/MS were acquired with a resolution of 35.000 and 15.000, respectively. An ion selection threshold of 150 counts per second (cps) was set. Probable TOF deviations were corrected by calibration before and within sample batches to ensure accuracy.

2.12 Metaproteomics Data Analysis

For label-free quantification, 300 wiff raw files were generated from LC-MS/MS representing the 300 injected slices, where each 10 slices represent one sample. Generated MS files were searched in ProteinPilot™ ((version 5.0.1), using Paragon search algorithm (SCIEX, USA), against the pseudo database generated by the 16S rRNA analysis. The false discovery rate (FDR) was set as 1% of the protein level, ensuring high-quality results. Trypsin was selected as a digestion factor and iodoacetamide was selected as

the Cys Alkylation.

2.13 Metaproteomics Statistical Analysis

Protein lists generated from the searches were exported from the search database and pre-processed using in-house software "ProteoCompanion" followed by "ProteoSelector" (https://www.57357.org/en/department/proteomics-unit-dept/in-house-bioinformatics-tools/) to merge the normalized spectral abundance factor (NSAF) into a single CSV file. After removing decoy hits, data were normalized using probabilistic quotient normalization (PQN) [33], log-transformed, and auto-scaled. Statistical analysis was proceeded using R coding, including protein filtering, by removing hits missed in >35% per group. Data were subjected to unpaired t-test, considering only hits with p-value < 0.05 and FDR-adjusted p-value < 0.05. Principle Component Analysis (PCA) was tested to perform samples' classification.

2.14 Gene Ontology-based Functional Analysis

Gene Ontology (GO) annotation for the significant differentially expressed proteins (DEPs) was retrieved with respect to Biological Process, Molecular Function, and Cellular Component ontologies, using UniProtR [34]. GO was done for the upregulated and downregulated proteins separately. Subsequently, GO-Biological Process annotations were analyzed by REVIGO [35] that used a clustering algorithm to summarize the list of GO terms and find a representative of the terms related to each other.

Chapter 3

Results

3.1 Microbial Profile of Patients with ALL

Stool samples were collected prospectively from 30 newly diagnosed ALL pediatric patients admitted at the CCHE-57357, following the inclusion and exclusion enrollment criteria. Stool samples were also collected from 30 healthy children following the same criteria. Samples were divided for subsequent metagenomics and metaproteomics processing. At first, following the metagenomics pipeline, DNA was extracted from the samples for 16S rRNA analysis and sequencing using Illumina MiSeq System. Generated raw reads were filtered to assure high quality reads before the downstream analysis. Reads length were greater than or equal to 300, and reads number were greater than 0.2 Mb. Regarding the ALL samples, two samples out of the 30, unfortunately, did not pass the QC criteria, and only 28 samples were used for the downstream analysis. Reads were further processed by the ASV error correction method of DADA2 platform using QIIME2 to trim adapter sequences and filter chimeric sequences. Exact number of filtered sequences in each sample can be found in Supplementary Table 1.

In the ALL samples, 3,221 non-redundant filtered features were identified, with a minimum length of 286 base pairs (bp) and a maximum length of 467 bp. The average length across samples was 449.66 bp. The mean frequency per sample was 41,457 (Supplementary Table 2, and Supplementary Figure 2).

The overall identified bacterial taxa were mainly branched from the domain bacteria. To a lesser extent, archaea was also identified, and was represented by the genus *Methanobrevibacter*, of the class *Methanobacteria*, from the *Euryarchaeota* phylum (Figure 1).

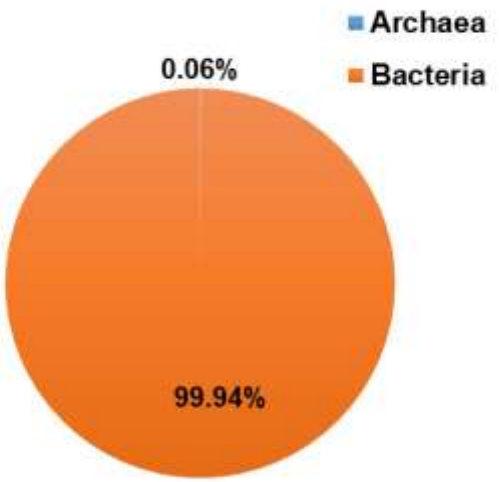


Figure 1: Microbiota composition of the ALL samples on the domain level. The percentage of each domain is shown.

Bacteria were represented by the following phyla, *Firmicutes*, *Proteobacteria*, *Actinobacteriota*, *Bacteroidota*, *Verrucomicrobiota*, *Desulfobacterota*, and *Synergistota*. The *Firmicutes* was the dominant phylum accounting for about 70% of the taxonomic composition, followed by *Bacteroidetes* with ~12% (Figure 2).

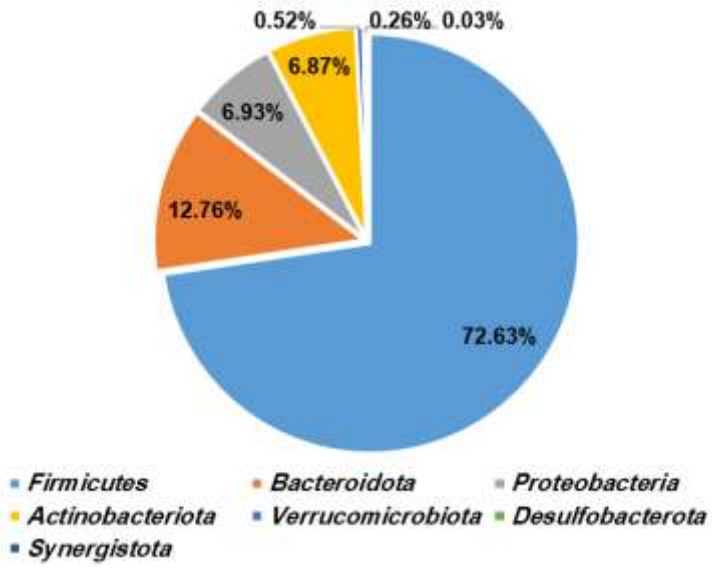


Figure 2: Microbiota composition of the ALL samples on the phylum level. The percentage of each phylum is shown

The abundance of *Proteobacteria* and *Actinobacteriota* was almost the same, 6% each. Other phyla were minimally presented in the ALL taxonomic composition, accounting for less than 1%, including *Verrucomicrobiota*, *Desulfobacterota*, and *Synergistota*. The relative abundance of the identified microbial composition in each ALL sample was shown in the bar plot (Figure 3).

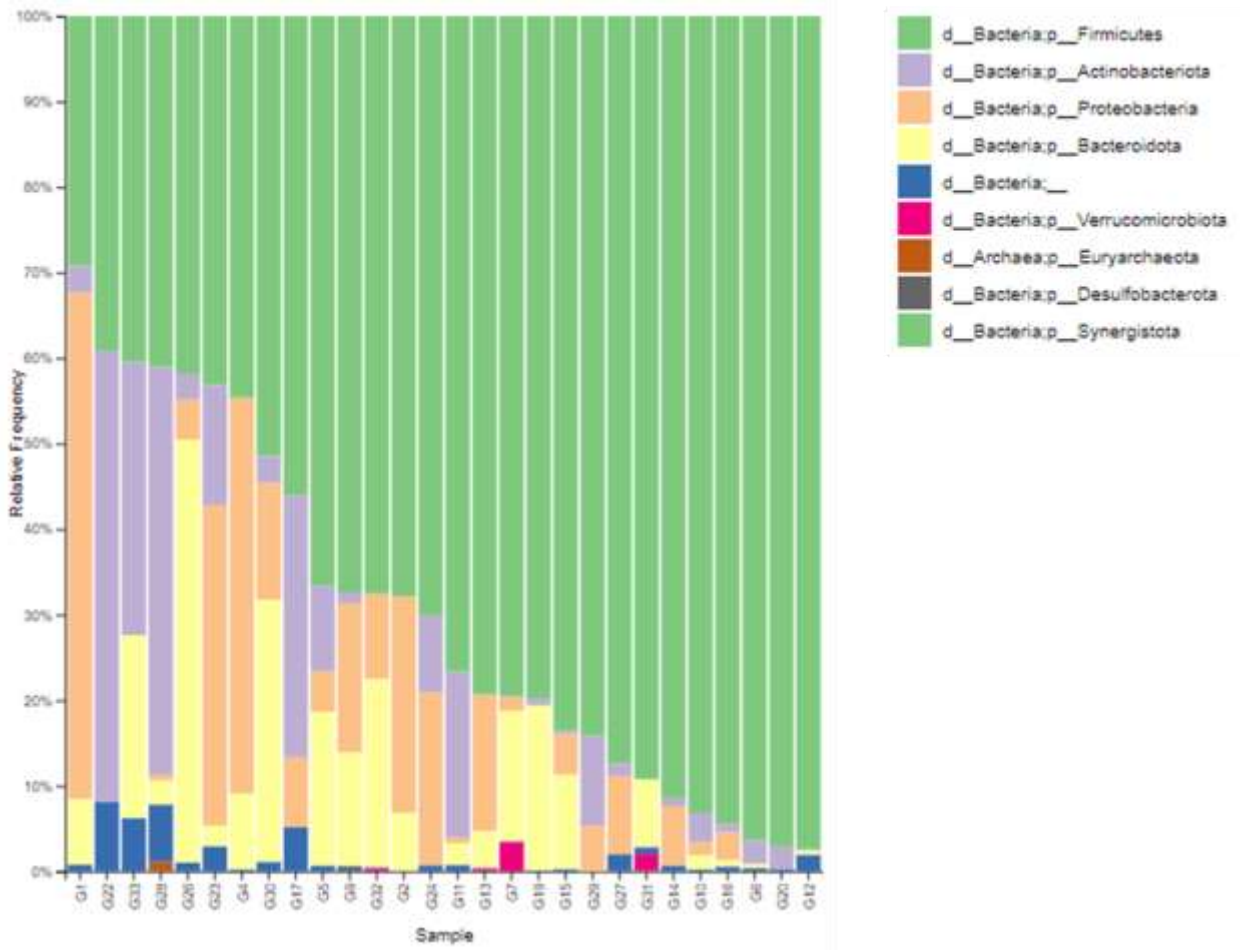


Figure 3: The relative abundance of the identified microbial composition in each ALL sample. The x-axis represents the samples, and the y-axis shows the relative frequency in percentage. The color legend of the phyla is shown.

The dominant *Firmicutes* phylum was characterized by the classes *Clostridia* (66.6%), *Bacilli* (26.3%), and *Negativicutes* (3.7%) (Figure 4). The remaining percentage unfortunately could not be attributed to lower taxonomic ranks.

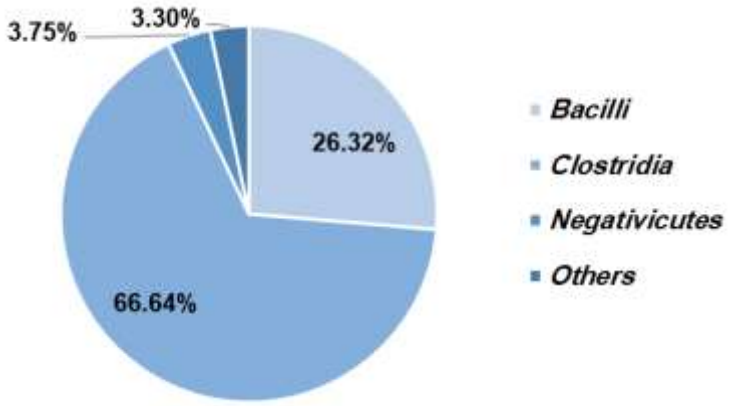


Figure 4: The most prevalent classes of the *Firmicutes* phylum in the ALL samples. The percentage of each class is shown, calculated from the total *Firmicutes*.

All the *Bacteroidetes* phylum was represented by the *Bacteroidia* class. The whole taxonomic profile of the ALL samples at the class level was shown in Figure 5.

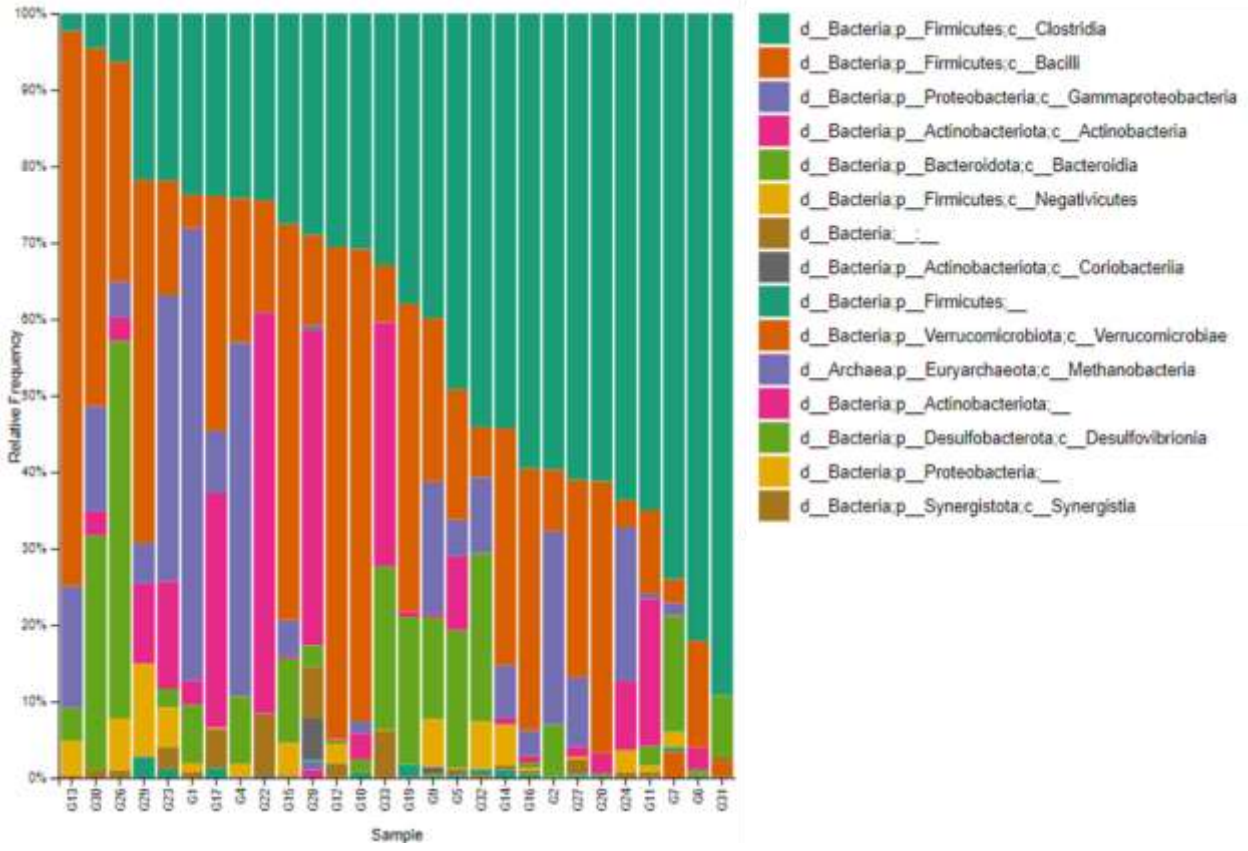


Figure 5: The relative abundance of the identified microbial composition, at the class level, in each ALL sample. The x-axis represents the samples, and the y-axis shows the relative frequency in percentage. The color legend of the phyla is shown.

A total of 115 non redundant genera were identified in the ALL pool. The most dominant genera from the *Firmicutes* phylum were *Streptococcus* (~7.5%), *Faecalibacterium* (7%), *Ruminococcus* (5.5%), followed by *Subdoligranulum* (3.8%), and *Eubacterium* (3.4%). Other less abundant genera from the *Firmicutes* were *Blautia, Clostridium*, *Lachnospiraceae*, and *Roseburia*. Regarding the *Bacteroidota* phylum, the most abundant genera were *Bacteroides*, *Parabacteroides* and *Prevotella*.

Principle component analysis of gut microbiota profile in controls and ALL patients revealed a strong group separation (Figure 6).

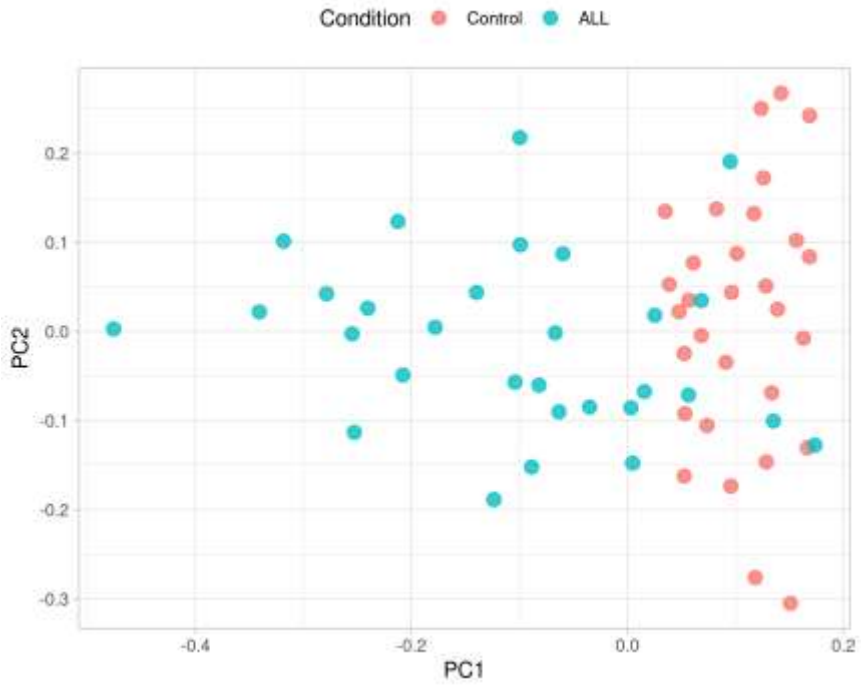


Figure 6: Principle component analysis of gut microbiota profile in the study cohorts.

The bar plot (Figure 7) shows the overall phyla composition of the control individuals compared to the ALL patients.

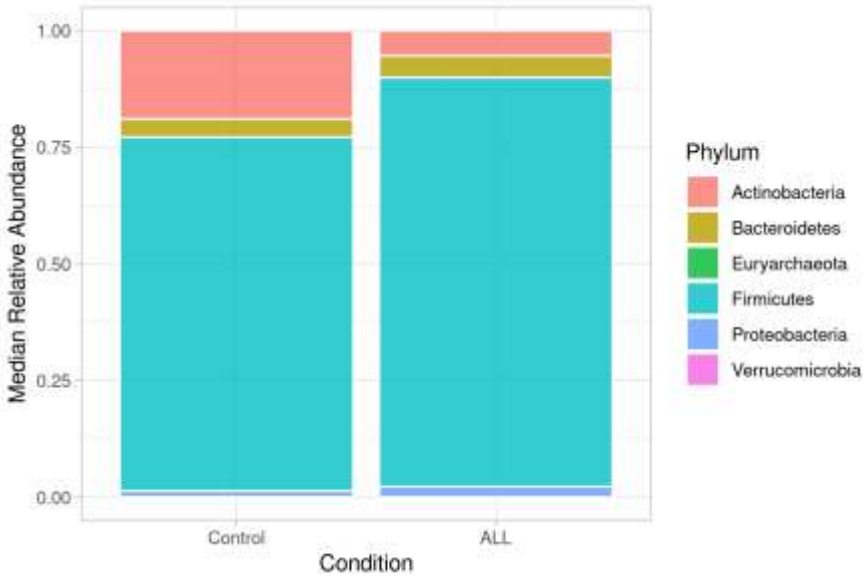


Figure 7: Microbiota Composition at phyla level, in controls and ALL patients. The x-axis represents the median relative abundance.

Differential analysis revealed significant differentially abundant genera between the control individuals and ALL patients, such as the higher abundance of *Streptococcus*, and the lower abundance of *Dialister*, *Lachnospira* (Figure 8).

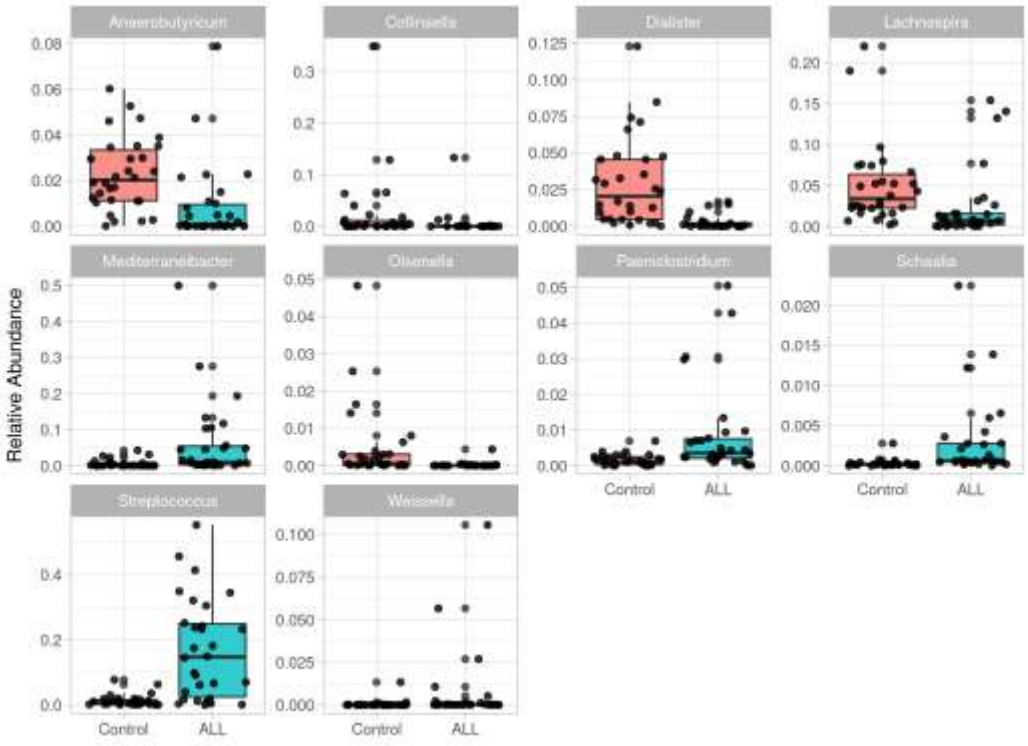


Figure 8: Differentially abundant genera between the controls and ALL patients.

3.2 Microbial Proteins Revealed by Metaproteomics

Shotgun proteomics analysis for 30 stool samples (15 from ALL patients and 15 from control individuals) was performed to characterize potential proteins. A total of 4495 proteins were identified and shared between the ALL and control (1664 after normalization and filtration discussed in Materials and methods section). Additionally, a total of 85 proteins were unique to either the ALL or control group. Shared proteins were processed for subsequent statistical comparison. The PCA analysis coherently separated significantly the 2 cohorts, with 46.6% discrimination using component 1 (Figure 9).

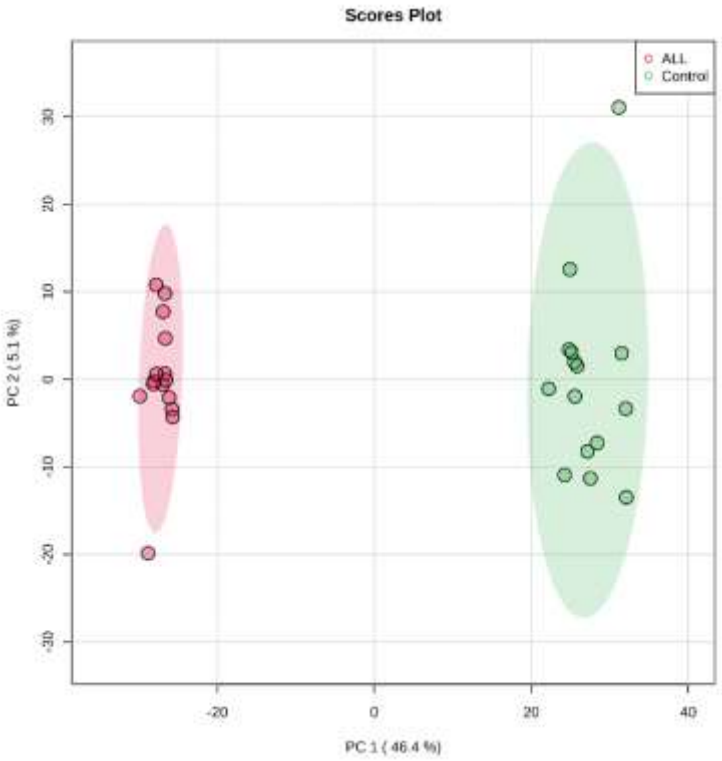


Figure 9: 2D principal component analysis (PCA) plot of the study cohorts, ALL and controls. PC1 and PC2 are the first and the second principal component, respectively.

This analysis revealed 1080 significantly differentially expressed proteins between ALL and control individuals (p-value < 0.05 and FDR-adjusted p-value < 0.05, fold change $\geq \pm 2$). Heat map analysis displayed the top 100 proteins' expression differences among the 2 cohorts (Figure 10), which showed complete clustering of the cohorts.

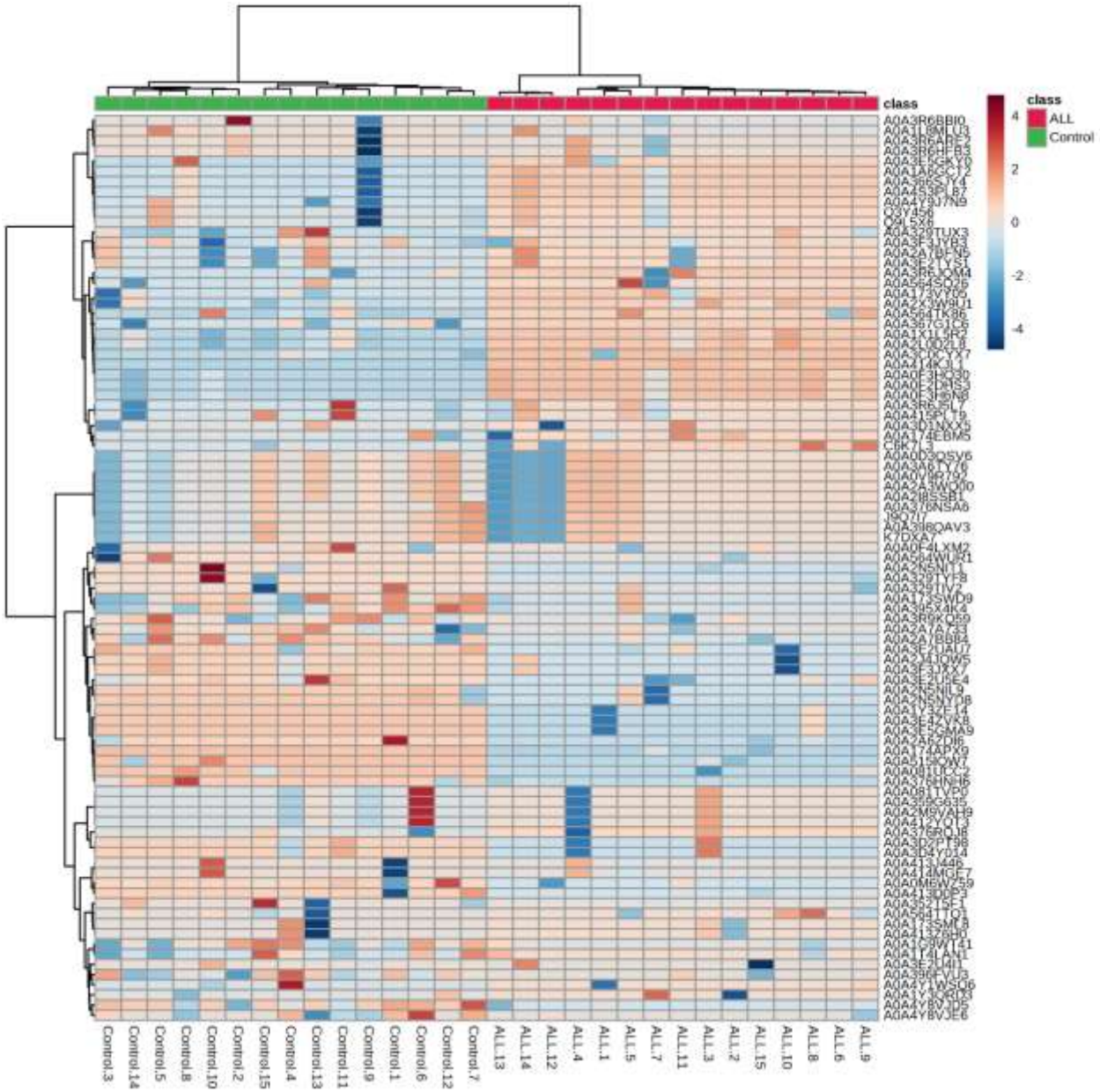


Figure 10: Hierarchical clustering of study samples showing the expression pattern of the top 100 significant proteins between ALL, and control. Significant proteins are shown on the y-axis, study cohorts are shown on the x-axis, displaying clusters of samples to study groups. Relative expression of each protein shown based on the z-score of the protein's normalized NSAF, with overexpressed and downregulated proteins depicted in red and blue, respectively.

Volcano plot revealed up-regulation of 414 proteins in ALL with a magnitude of ≥2-fold change (Figure 8). On the other hand, 667 proteins were significantly down-regulated in ALL (FDR-adjusted p-value < 0.1, fold change ≤ -2) (Figure 11).

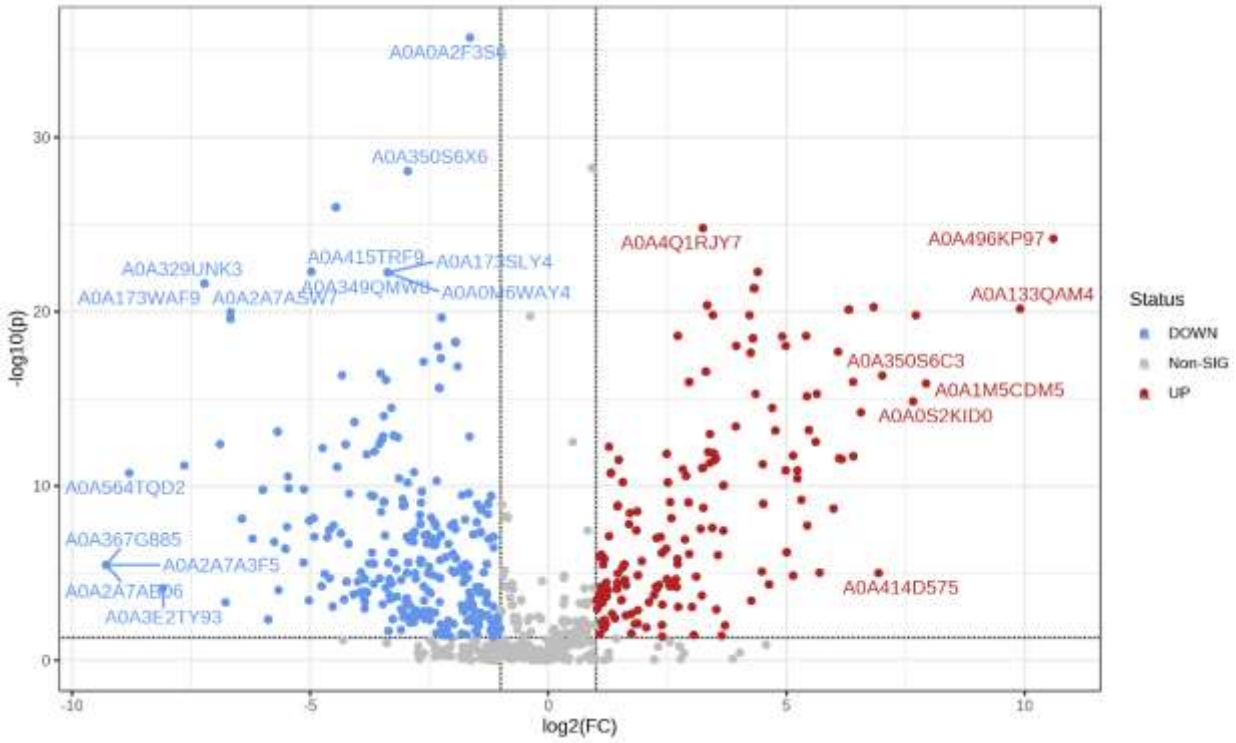


Figure 11: Volcano plot representing expression profile of proteins that differ significantly between pediatric ALL patients and control individuals. Horizontal grey line represents $-\log_{10}$(adjusted pvalue) of 1.3. Vertical grey lines represent a magnitude of either ±1 $\log_2$ fold change. Red and blue dots represent significantly differentially expressed proteins (p-value < 0.05, FDR-adjusted p-value < 0.05 and $\log_2$ fold change ≥ ±1). Non-significant proteins are shown as grey dots.

Due to the huge microbial heterogeneity, the same protein could be secreted from different bacterial species, and hence it got different accession number. To overcome such issue, the analysis of differentially expressed proteins was done aside of its microbial source. With this notion, the significantly upregulated proteins resulted in 79 non-redundant proteins, including ABC transporters, Bacterial extracellular solute-binding protein, Extracellular solute-binding protein, Flagellin, Glyceraldehyde-3-phosphate

dehydrogenase, Ion-translocating oxidoreductase complex subunit C, Lipoprotein, Major capsid protein, Malate dehydrogenase, Rubrerythrin, TonB-dependent receptor, NADH peroxidase, OmpA family protein, OmpC, Outer membrane protein C, RagB/SusD family nutrient uptake outer membrane protein, Sporulation protein, SusD family protein, Xylose isomerase. Considering the downregulated 667 significant proteins, they resulted in 116 non-redundant proteins, including Aldose 1-epimerase, Branched chain amino acid aminotransferase, Lipoprotein, Malate dehydrogenase, OmpC, Outer membrane porin OmpF, SusD/RagB family nutrient-binding outer membrane lipoprotein, and Xylose isomerase.

Interestingly, 58 proteins were found unique to the ALL group, including Alcohol dehydrogenase, FAD-binding protein, FAD-linked oxidoreductase, LacI family transcriptional regulator, SDR family NAD(P)-dependent oxidoreductase, and Thioredoxin. On the other hand, 27 proteins were unique to the control group, such as Acetoacetate decarboxylase family protein, Polygalacturonase, and Carbon starvation protein.

3.3 Functional Assignment of the Gut Microbial Proteins in Patients with ALL

To further depict the biological functions of the differentially expressed proteins in ALL patients. Gene Ontology (GO) annotation was done for the upregulated and downregulated proteins separately. Proteins were annotated with respect to Biological Process, Molecular Function, and Cellular Component ontologies. As GO Biological Process, significant upregulated proteins in ALL were enormously related to pectin Catabolic Process, chemotaxis, signal Transduction, phosphorylation, glycolytic process and glucose metabolic process, with the last being the most frequent (Figure 12). On the Molecular Function level, upregulated proteins were massively annotated to NAD and NADP binding, and oxidoreductase activity, in addition to iron ion binding, glucuronate isomerase activity, peroxidase activity, ATP binding, and L-malate dehydrogenase activity. Most proteins were annotated with the cellular component terms including cytoplasm, membrane, cell outer membrane, and bacterial-type flagellum.

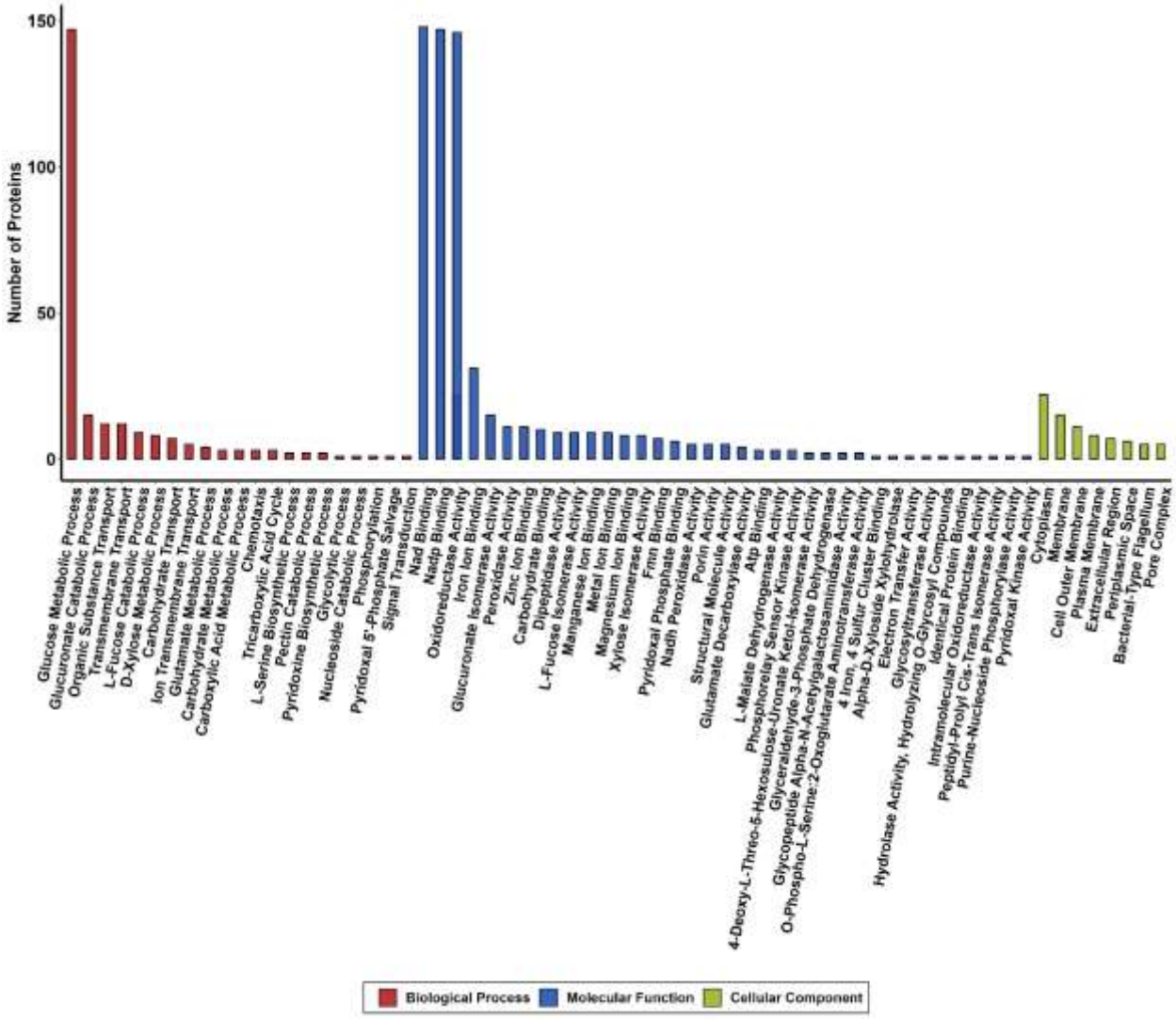


Figure 12: Gene Ontology (GO) annotations of the significant upregulated proteins in ALL compared to control. GO annotations, according to the Biological Process, Molecular Function, and Cellular Localization, are shown on the x-axis, whereas the y-axis represents the number of proteins.

Analysis of the Biological Process GO using the Web server REVIGO revealed several clustered pathways. In other words, the biological processes, to which the upregulated proteins were annotated, can be grouped to several clusters represented by a representative to reduce redundancy and ease the interpretation (Figure 13). Upregulated proteins in ALL patients were, generally, related to pectin catabolic process, pyridoxine

biosynthetic process, and chemotaxis. Furthermore, some GO terms did not show any similarity with others, and hence, they showed no clustering, including tricarboxylic acid cycle, and phosphorylation.

Figure 13: TreeMap visualization of REVIGO showing the GO Biological Process annotations of the upregulated proteins in ALL. Rectangles with the same color represents one cluster.

GO analysis of unique proteins to ALL revealed annotations related to iron import into cell, oxidoreductase activity, glucose metabolic process, and pectin catabolic process, in alignment with the observed functions in the differentially upregulated proteins in ALL. Similarly, the results of the REVIGO analysis of the GO annotations of the unique proteins in controls were aligned with the observed functions.

Regarding the downregulated proteins in ALL patients, proteins were mostly annotated to Ion Transmembrane Transport, Leucine and valine biosynthetic process, carbohydrate metabolic process, butyrate biosynthetic process, and acetate metabolic process (as GO Biological Process) (Figure 14-15). On the Molecular Function level, porin activity, and L-Leucine, L-Isoleucine, and L-valine transaminase Activity were the most annotations. Proteins were annotated to cell outer membrane, pore complex, and cytoplasm, as cellular component terms.

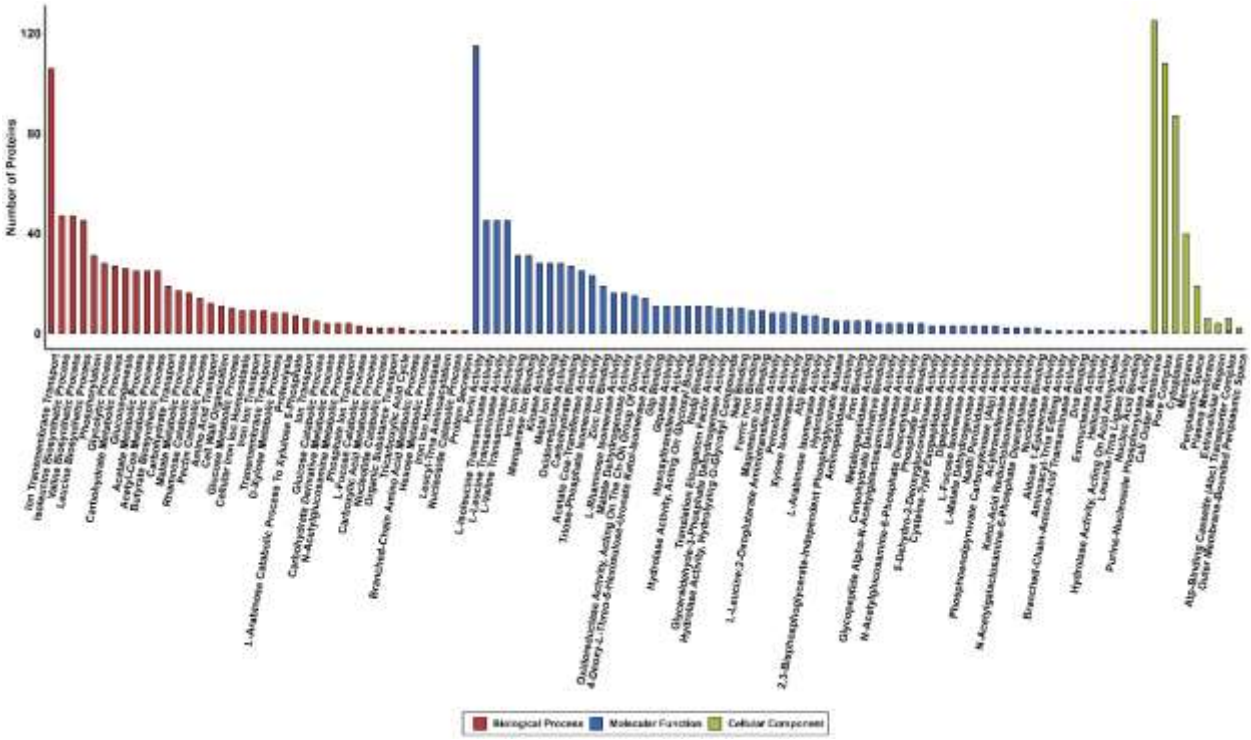


Figure 14: Gene Ontology (GO) annotations of the significant downregulated proteins in ALL compared to control. GO annotations, according to the Biological Process, Molecular Function, and Cellular Localization, are shown on the x-axis, whereas the y-axis represents the number of proteins.

Figure 15: TreeMap visualization of REVIGO showing the GO Biological Process annotations of the downregulated proteins in ALL patients. Rectangles with the same color represents one cluster.

Chapter 4

Discussion

The microbiota has been recently linked to a variety of disease status, including cancer, and hence, studying the microbiota in cancer opened the door to new era of cancer strategies to further understand the disease pathophysiology. In this vein, stool samples were collected from pediatric ALL patients for multi-omics analysis. At first, the stool samples underwent 16s rRNA analysis using the Illumina MiSeq System. After applying filtration criteria, our results revealed the predominance of *Firmicutes* in ALL patients, accounting for more than 70%, which was consistent with previous ALL metagenomics study [36]. However, another metagenomics study has reported the prevalence of *Bacteroidetes* over the *Firmicutes* in ALL [37]. Such discrepancy further highlights the importance of studying the *Firmicutes* phylum and its changes in ALL. It worth noting that other factors contribute to changes in the microbial compositions including the diet, and geographical regions [3], which could be a possible explanation for the discrepancy seen in *Firmicutes* across the different cohorts. In fact, the *Firmicutes* phylum plays a crucial role in the digestion, especially breaking down the carbohydrates and polysaccharides to SCFAs [38], which might indicate its susceptibility to dietary changes. Furthermore, the high abundance of the genera *Streptococcus* in ALL patients might indicate an inflammatory increase as those bacteria stimulate the production of cytokines including IL-6, IL-8, and IL-1β [39].

The prevalence of *Bacteroides* and *Prevotella* in ALL, as shown by our results, comes with agreement with previous ALL metagenomics studies, where the abundance of *Bacteroides* specifically was suggested as potential marker for the dysbiosis status in ALL [24, 37]. In fact, both genera play a role in the gut immunomodulation through induction of IL-17 in the gut mucosa via T helper 17 (TH17) cells, which further increase the inflammation [40]. The high abundance of the *Parabacteroides* was previously reported in chronic lymphocytic leukemia [41], however its potential implication in cancer remains unclear.

Interestingly, some genera showed lower abundance in the ALL samples such as *Blautia*. In fact, *Blautia* are commensal bacteria that produce SCFAs contributing to an anti-inflammatory function [42]. Therefore, our metageomics results indicate that ALL can cause inflammation by disrupting the anti-inflammatory processes. Other less prevalent genera in ALL patients were *Lachnospiraceae* and *Roseburia,* consistent with previous ALL metagenomics studies [24, 36]. In fact, both bacterium produce the SCFA, butyrate. The latter is crucial for maintaining the gut homeostasis and the integrity of the mucus layer, in addition to provide the energy required for the gut epithelial cells [24]. Furthermore, *Lachnospiraceae* and *Roseburia* protect the host from pathogenic bacteria by decreasing the intestinal lumen pH [36]. Moreover, both genera have robust anti-inflammatory functions. And hence, the lower presence of these genera in ALL patients pinpoint an increased mucositis and decreased epithelial integrity.

The ultimate goal of metaproteomics is to elucidate the functional characteristics associated with the changes in the microbial composition at dysbiosis. Unfortunately, the high heterogeneity and similarity of the proteins secreted from same and/or different bacteria still imposes some challenges on analyzing the proteins. To overcome, functional analysis of the proteins is done using functional databses such as GO, eggNOG, and KEGG. In our study, upregulated and downregulated proteins in ALL were annotated to GO. Regarding the upregulated proteins, they were found enormously related to iron binding. In fact, iron is essential for enteric pathogens to invade and colonize the gut and become more virulent [23]. Furthermore, our results found two main differential proteins involved in iron related functions, ABC transporters, TonB-dependent receptors, and Rubrerythrin. ABC transporters and TonB-dependent receptors regulate the concentration of iron in the intestinal lumen and participate in the transport of siderophores. The latter are small chelators with high iron affinity, produced by bacteria, mainly Gram-negative bacteria, to uptake the necessary iron for their growth [43]. Rubrerythrin, a non heme iron protein, is involved in the storage and detoxification of iron, in addition to oxidative stress defense [23]. Collectively, our results pinpointed an increased iron concentration and iron demand by the gut microbiota of ALL patients.

Furthermore, our results showed that some upregulated proteins were highly related to NAD and NADP

binding, and oxidoreductase activity, indicating potential higher oxidative stress levels in ALL. Increased oxidative stress has been observed before in gut microbiota of colorectal cancer patients by metaproteomics [23]. In fact, iron is directly linked to oxidative stress as it converts the less reactive hydrogen peroxide to hydroxyl radical and ferryl iron which are more reactive oxygen species, via Fenton reaction [44]. Increased oxidative stress can cause DNA damage of the gut epithelial cells.

Also, ALL patients showed enormous GO annotation to Glucose Metabolic Process. In fact, on one hand, gut microbiota play a role in regulating the glucose metabolism and homeostasis [45]. On the other hand, cancer cells need high amount of glucose to supply its energy needs [46]. In this vein, whether there was a potential relationship between the glucose metabolism by gut microbiota and energy requirement by ALL cancerous cells still require further studies. Another highly annotated GO in ALL was Pectin Catabolic Process. Pectins are fibers generated from polysaccharides of some plants cell wall, that cannot be digested by host enzymes, and are degraded by beneficial bacteria [47]. Pectins are known with their ability to maintain the gut epithelial integrity, and modulating immune functions. Studies on the exact effect of pectin on the immune system are controversy, some highlighted activation of dendritic cells and macrophages, while others mentioned inhibition [48]. And hence, the exact role of pectins-related proteins in ALL is still unclear.

Downregulated proteins in ALL patients were mostly annotated to carbohydrate metabolic process, and butyrate biosynthetic process. In fact, carbohydrates are crucial for maintaining the intestinal integrity, protecting from pathogenic bacteria, and reducing the inflammatory responses [49]. A decrease in carbohydrates metabolic functions was also observed in colon cancer patients [49].

Moreover, GO annotations associated with downregulated proteins were related to amino acids metabolism, especially L-Leucine, L-Isoleucine, and L-Valine. Such amino acids were found essential to some immune functions, including the enhancement of the $CD8^+$ T cell activity [50]. Furthermore, they could improve the anti-oxidative functions against oxidative stress.

The integration of the observations inferred from the two omics techniques is the key goal in an attempt to

understand the complete picture of microbial dysbiosis associated with ALL (Figure 16). For instance, the less abundance of SCFAs and butyrate-secreting bacteria, *Blautia*, *Lachnospiraceae*, *Roseburia*, observed by our microbiota profiling results is aligned with the observed decrease in proteins, and unique proteins in ALL, involved in butyrate biosynthetic and carbohydrate metabolic process, leading to decreased gut mucosal integrity and increased permeability and inflammation in ALL. Furthermore, the high frequency of certain immunomodulatory bacteria, *Streptococcus*, *Prevotella*, and *Bacteroides*, could be correlated with the downregulated amino acids metabolism functions on the metaproteomics level, contributing to further increased inflammation through cytokines release and weakened immunity (weaker $CD8^+$ T cell activity). Moreover, the functional characteristics of the unique proteins in the control individuals further reflected the microbiota profile under healthy status with multiple functions contributing to the ultimate homeostasis for the host benefit. On the other hand, the gut dysbiosis status reflected by the differential and unique proteins in ALL highlighted the potential increased the iron demand causing elevated oxidative stress, and hence, inducing DNA damage in the gut epithelial cells, which might further worsen the cancer status in ALL patients.

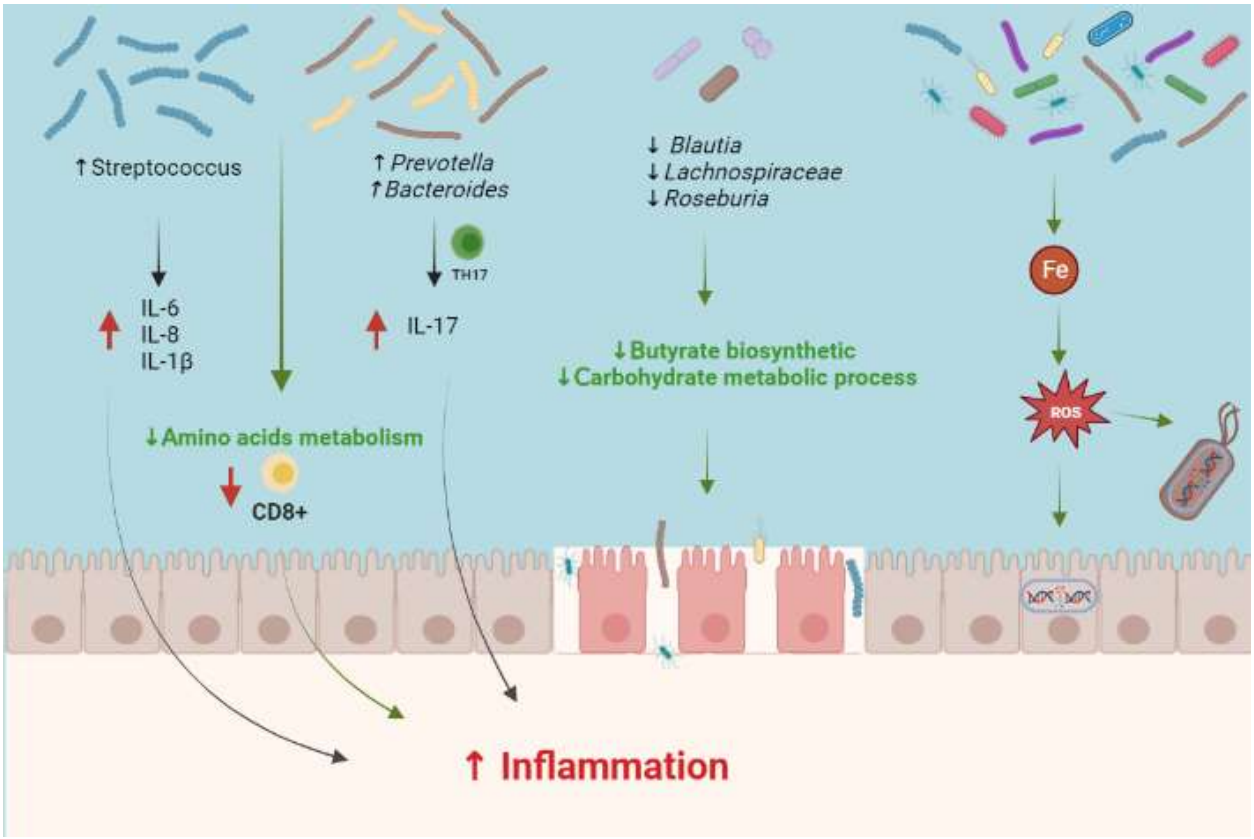


Figure 16. Gut microbiota in pediatric ALL patients

Chapter 5

Conclusion

In conclusion, this study explored the gut microbial functional characteristics in pediatric ALL patients. We have explored the microbial composition in ALL patients through 16sRNA microbial profiling, which revealed the high abundance of some genera as *Bacteroides*, *Prevotella*, and *Streptococcus*, compared to other less abundant organisms, such as *Blautia*, *Lachnospiraceae*, *Roseburia*. Our metaproteomics results inferred about the functional implications of ALL gut microbiota. We postulated higher iron demand and oxidative stress in ALL patients compared to healthy individuals. Furthermore, functions related to amino acids, carbohydrates and butyrate metabolism were downregulated in ALL. These promising results are a preliminary step toward a deeper understanding of the gut microbiota in ALL.

References

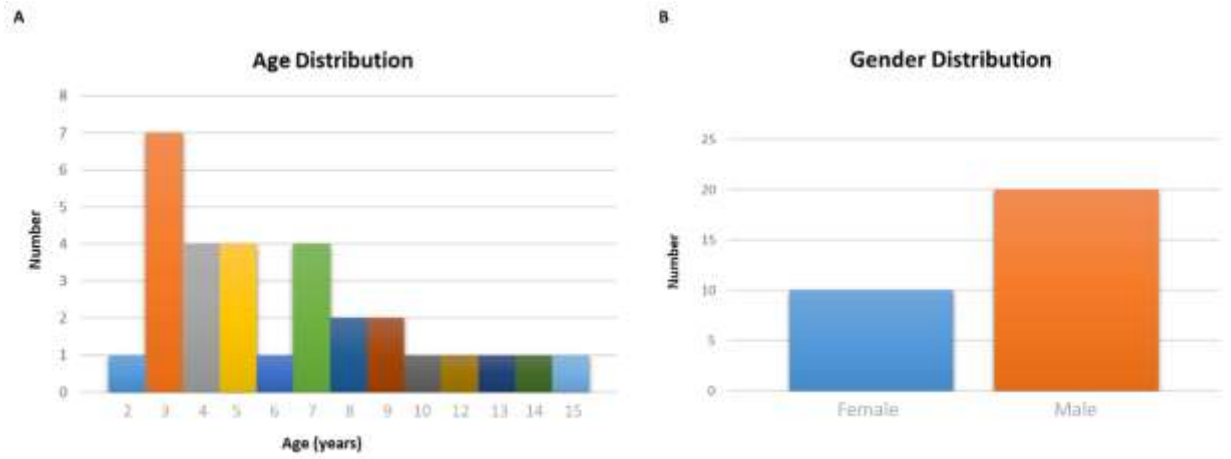
1. Ursell, L.K., et al., *Defining the human microbiome.* Nutrition Reviews, 2012. **70**(suppl_1): p. S38-S44.
2. Rinninella, E., P. Raoul, and M. Cintoni, *What is the Healthy Gut Microbiota Composition? A Changing Ecosystem across Age, Environment, Diet, and Diseases.* 2019. **7**(1).
3. Thursby, E. and N. Juge, *Introduction to the human gut microbiota.* Biochem J, 2017. **474**(11): p. 1823-1836.
4. Gilbert, J.A., et al., *Current understanding of the human microbiome.* Nature Medicine, 2018. **24**(4): p. 392-400.
5. Ezzeldin, S., et al., *Current Understanding of Human Metaproteome Association and Modulation.* 2019. **18**(10): p. 3539-3554.
6. Tan, J., et al., *Chapter Three - The Role of Short-Chain Fatty Acids in Health and Disease*, in *Advances in Immunology*, F.W. Alt, Editor. 2014, Academic Press. p. 91-119.
7. Bhatt, A.P., M.R. Redinbo, and S.J. Bultman, *The role of the microbiome in cancer development and therapy.* CA Cancer J Clin, 2017. **67**(4): p. 326-344.
8. Wang, W.L., et al., *Application of metagenomics in the human gut microbiome.* World J Gastroenterol, 2015. **21**(3): p. 803-14.
9. Savage, D.C., *Microbial ecology of the gastrointestinal tract.* Annu Rev Microbiol, 1977. **31**: p. 107-33.
10. Methé, B.A., et al., *A framework for human microbiome research.* Nature, 2012. **486**(7402): p. 215-221.
11. Qin, J., et al., *A human gut microbial gene catalogue established by metagenomic sequencing.* Nature, 2010. **464**(7285): p. 59-65.
12. Li, J., et al., *An integrated catalog of reference genes in the human gut microbiome.* Nature Biotechnology, 2014. **32**(8): p. 834-841.
13. Ferrer, M., et al., *Microbiota from the distal guts of lean and obese adolescents exhibit partial functional redundancy besides clear differences in community structure.* Environ Microbiol, 2013. **15**(1): p. 211-26.
14. Erickson, A.R., et al., *Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease.* PLoS One, 2012. **7**(11): p. e49138.
15. Forslund, K., et al., *Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota.* Nature, 2015. **528**(7581): p. 262-266.
16. Thomas, A.M., et al., *Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation.* Nature Medicine, 2019. **25**(4): p. 667-678.
17. Bhatt, A.P., M.R. Redinbo, and S.J. Bultman, *The role of the microbiome in cancer development and therapy.* CA: a cancer journal for clinicians, 2017. **67**(4): p. 326-344.
18. Dai, Z., et al., *Multi-cohort analysis of colorectal cancer metagenome identified altered bacteria across populations and universal bacterial markers.* Microbiome, 2018. **6**(1): p. 70.
19. Jain, M. and R. Yadav, *Metagenomics Analysis of Breast Cancer to Study Bacterial Diversity.* Journal of Drug Delivery and Therapeutics, 2022. **12**(2-S): p. 119-126.
20. Liu, Y., et al., *Metagenomic Analysis Reveals a Changing Microbiome Associated With the Depth*

of Invasion of Oral Squamous Cell Carcinoma. Frontiers in microbiology, 2022. **13**.
21. Petriz, B.A. and O.L. Franco, *Metaproteomics as a Complementary Approach to Gut Microbiota in Health and Disease.* Front Chem, 2017. **5**: p. 4.
22. Lee, P.Y., et al., *Metaproteomic analysis of human gut microbiota: where are we heading?* Journal of Biomedical Science, 2017. **24**(1): p. 36.
23. Long, S., et al., *Metaproteomics characterizes human gut microbiome function in colorectal cancer.* npj Biofilms and Microbiomes, 2020. **6**(1): p. 14.
24. Rajagopala, S.V., et al., *Gastrointestinal microbial populations can distinguish pediatric and adolescent Acute Lymphoblastic Leukemia (ALL) at the time of disease diagnosis.* BMC Genomics, 2016. **17**(1): p. 635.
25. Bai, L., et al., *Changes in the gastrointestinal microbiota of children with acute lymphoblastic leukaemia and its association with antibiotics in the short term.* J Med Microbiol, 2017. **66**(9): p. 1297-1307.
26. Ward, E., et al., *Childhood and adolescent cancer statistics, 2014.* CA: a cancer journal for clinicians, 2014. **64**(2): p. 83-103.
27. Carroll, W.L. and E.A. Raetz, *Clinical and Laboratory Biology of Childhood Acute Lymphoblastic Leukemia.* The Journal of Pediatrics, 2012. **160**(1): p. 10-18.
28. Klindworth, A., et al., *Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies.* Nucleic Acids Res, 2013. **41**(1): p. e1.
29. Bolyen, E., et al., *Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2.* Nature Biotechnology, 2019. **37**(8): p. 852-857.
30. Shannon, C.E., *A mathematical theory of communication.* The Bell system technical journal, 1948. **27**(3): p. 379-423.
31. Xiong, W., et al., *Microbial metaproteomics for characterizing the range of metabolic functions and activities of human gut microbiota.* Proteomics, 2015. **15**(20): p. 3424-3438.
32. Smith, P.K., et al., *Measurement of protein using bicinchoninic acid.* Anal Biochem, 1985. **150**(1): p. 76-85.
33. Dieterle, F., et al., *Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics.* Anal Chem, 2006. **78**(13): p. 4281-90.
34. Soudy, M., et al., *UniprotR: Retrieving and visualizing protein sequence and functional information from Universal Protein Resource (UniProt knowledgebase).* J Proteomics, 2020. **213**: p. 103613.
35. Supek, F., et al., *REVIGO summarizes and visualizes long lists of gene ontology terms.* PLoS One, 2011. **6**(7): p. e21800.
36. Bai, L., et al., *Changes in the gastrointestinal microbiota of children with acute lymphoblastic leukaemia and its association with antibiotics in the short term.* Journal of Medical Microbiology, 2017. **66**(9): p. 1297-1307.
37. Chua, L.L., et al., *Temporal changes in gut microbiota profile in children with acute lymphoblastic leukemia prior to commencement-, during-, and post-cessation of chemotherapy.* 2020. **20**(1): p. 151.
38. Houtman, T.A., et al., *Gut microbiota and BMI throughout childhood: the role of firmicutes, bacteroidetes, and short-chain fatty acid producers.* Scientific Reports, 2022. **12**(1): p. 3140.
39. Ellmerich, S., et al., *Promotion of intestinal carcinogenesis by Streptococcus bovis.*

Carcinogenesis, 2000. **21**(4): p. 753-6.
40. Gao, Z., et al., *Microbiota disbiosis is associated with colorectal cancer.* Front Microbiol, 2015. **6**: p. 20.
41. Faitová, T., et al., *The gut microbiome in patients with chronic lymphocytic leukemia.* Haematologica, 2022. **107**(9): p. 2238-2243.
42. Oldenburg, M., et al., *The Microbiome in Childhood Acute Lymphoblastic Leukemia.* Cancers, 2021. **13**(19): p. 4947.
43. Schalk, I.J. and L. Guillon, *Fate of ferrisiderophores after import across bacterial outer membranes: different iron release strategies are observed in the cytoplasm or periplasm depending on the siderophore pathways.* Amino Acids, 2013. **44**(5): p. 1267-1277.
44. Touati, D., *Iron and oxidative stress in bacteria.* Arch Biochem Biophys, 2000. **373**(1): p. 1-6.
45. Portincasa, P., et al., *Gut Microbiota and Short Chain Fatty Acids: Implications in Glucose Homeostasis.* International Journal of Molecular Sciences, 2022. **23**(3): p. 1105.
46. Cazzaniga, M., et al., *Gut Microbiota, Metabolic Disorders and Breast Cancer: Could Berberine Turn Out to Be a Transversal Nutraceutical Tool? A Narrative Analysis.* International Journal of Molecular Sciences, 2022. **23**(20): p. 12538.
47. Larsen, N., et al., *Potential of Pectins to Beneficially Modulate the Gut Microbiota Depends on Their Structural Properties.* Front Microbiol, 2019. **10**: p. 223.
48. Beukema, M., M.M. Faas, and P. de Vos, *The effects of different dietary fiber pectin structures on the gastrointestinal immune barrier: impact via gut microbiota and direct effects on immune cells.* Exp Mol Med, 2020. **52**(9): p. 1364-1376.
49. Cheng, X., et al., *A review: Roles of carbohydrates in human diseases through regulation of imbalanced intestinal microbiota.* Journal of Functional Foods, 2020. **74**: p. 104197.
50. Liu, B., et al., *Exploring Gut Microbiome in Predicting the Efficacy of Immunotherapy in Non-Small Cell Lung Cancer.* Cancers, 2022. **14**(21): p. 5401.

Appendix 1

Supplementary Data



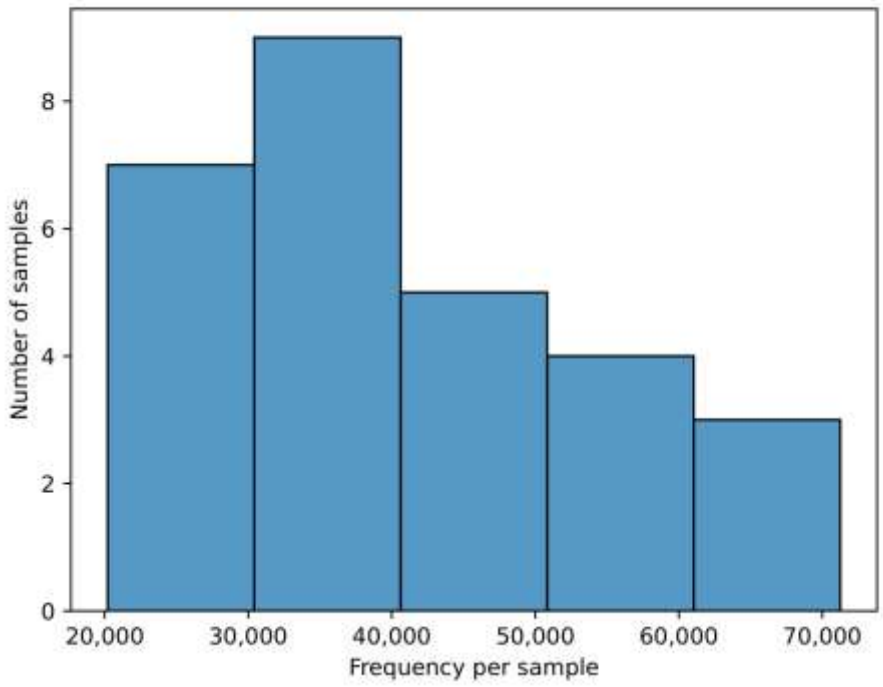
Supplementary Figure 1: Study samples description. (A) Patients' age distribution. (B) Gender distribution.

Supplementary Table 1: Results of ASV error correction with DADA2 for sequences filtering

Sample-id	Input	Filtered	% of input passed filter	Denoised	Merged	% of input merged	Non-chimeric	% of input non-chimeric
G1	322590	239655	74.29	231219	161184	49.97	51251	15.89
G10	404729	291742	72.08	275231	206278	50.97	53666	13.26
G11	403155	290695	72.11	282528	233176	57.84	43969	10.91
G12	270731	193065	71.31	189209	161241	59.56	47997	17.73
G13	396862	276798	69.75	272156	230296	58.03	64961	16.37
G14	198483	142730	71.91	133427	88326	44.5	23009	11.59
G15	184286	134427	72.94	123519	75375	40.9	20210	10.97
G16	205278	154956	75.49	143212	87575	42.66	23369	11.38
G17	394861	277747	70.34	266426	195345	49.47	51965	13.16
G19	520180	407653	78.37	397590	322157	61.93	30415	5.85
G2	318014	245728	77.27	237723	177794	55.91	35682	11.22
G20	464156	355759	76.65	338339	209308	45.09	37789	8.14
G22	514323	354976	69.02	346980	268774	52.26	65981	12.83
G23	460011	332909	72.37	321112	254939	55.42	60978	13.26
G24	447127	334433	74.8	317742	223057	49.89	33946	7.59
G26	620887	433751	69.86	417063	312268	50.29	32705	5.27
G27	449287	324183	72.15	307672	234971	52.3	49999	11.13
G28	484218	317635	65.6	289266	192113	39.67	71254	14.72
G29	191167	130496	68.26	120438	65518	34.27	28383	14.85
G30	258114	186126	72.11	181421	143056	55.42	23920	9.27
G31	344358	243425	70.69	204654	103925	30.18	38943	11.31
G32	290409	222199	76.51	202386	116333	40.06	23893	8.23
G33	324458	238061	73.37	224845	147620	45.5	46726	14.4
G4	345873	261608	75.64	252904	187367	54.17	39377	11.38
G5	258881	193269	74.66	173376	103010	39.79	39187	15.14
G6	325820	247138	75.85	233935	158945	48.78	39584	12.15
G7	427195	317856	74.41	287703	141474	33.12	31699	7.42
G9	509709	365628	71.73	347699	229156	44.96	49957	9.8

Supplementary Table 2: Mean reads frequency per sample

Min. frequency	1st Quartile	Median Frequency	3rd Quartile	Max. Frequency	Mean Frequency
20,210	31,378	39,282	50,312	71,254	41,457.67



Supplementary Figure 2: Overall distribution of reads frequency in each ALL sample.