UNIVERSITÀ DEGLI STUDI DI NAPOLI
"FEDERICO II"

DOTTORATO DI RICERCA
IN
Bioinformatica e biologia computazionale

XXVI CICLO

TESI DI DOTTORATO

"Next generation genomic sequencing and bioinformatics approaches for the study of the human gut microbiome in selected diseases of the human gut"

Coordinatore del Dottorato

Ch.mo Prof.
Sergio Cocozza

Dottorando

Giorgio Casaburi

Tutor

Ch.mo Prof.
Francesco Salvatore

ANNO ACCADEMICO 2013-2014

UNIVERSITÀ DEGLI STUDI DI NAPOLI
"FEDERICO II"

DOCTORAL THESIS

"Next generation genomic sequencing and bioinformatics approaches for the study of the human gut microbiome in selected diseases of the human gut"

GIORGIO CASABURI

MARCH 2014

INDEX

ABSTRACT

The advent of next generation sequencing technologies (NGS) has deeply transformed today's biology. Thanks to this new approach, many areas of study have been developed and scientists have the ability to analyze nucleic acids of any biological entity. Metagenomics is one of the most recent and promising fields among NGS applications. It allows microbial community analysis within a specific environment in order to obtain knowledge on genomes and taxonomic composition of environmental microbes and entire microbial communities. In this context, the human microbiota, represented by the total ecological community of commensal, symbiotic, and pathogenic microorganisms coexisting in our bodies (Lederberg J; *Scientist*. 2001;15:8), is drawing research's attention as it plays a central role in maintaining healthy status or leading to disease conditions (Wang, Zi-Kai *WJG*. 2013:1541). Particularly, the human gut is colonized by thousands of different microbial species, of which several billions are bacteria involved in important functions: gut permeability, immune system development and activation, metabolic function and colonization resistance (Prakash S. *Nature*. 2012;231-241). Many evidences relate the role of the gut microbiome in common bowel inflammatory diseases and autoimmune disorders (Xavier R. J. *Nature*. 2007:427-434).

Here, I present two distinct studies based on amplicon metagenomics analysis in Crohn's disease and celiac disease, using next generation sequencing techniques. The aim of the project is to deeply analyze the gut microbiome composition, from duodenal biopsies of Crohn's disease and celiac disease affected patients. The first study compares the microbiome bacterial composition of a child affected by Crohn's disease before and after a nutritional therapy together with a matched healthy control. The second study pertains to the gut microbiome characterization (bacteria and fungi) in adults with active celiac disease, compared with healthy controls and a group of non-active celiac disease patients, following a gluten free diet. The goal is the identification of microbial signatures that could be related to the pathogenesis or contribute to the disease phenotype in both Crohn's and celiac disease.

CHAPTER 1

INTRODUCTION

1.1 Next generation sequencing technologies

Deciphering nucleic acid sequences has always been a major interest among scientific community. The early 21st century was characterized by the announcement, for the first time in the human history, of the complete decryption of the human genetic code [1]. Thanks to the use of capillary electrophoresis (CE)-based Sanger sequencing, scientists gained the ability to reach paramount steps in biomedical research. Consequently, Sanger technology was adopted in many laboratories around the world. Besides the new promising results obtained and the important contribution made to biomedical research and molecular diagnostics, Sanger sequencing has always been confined by inherent limitations in throughput, versatility, speed, resolution and cost. Those limits often preclude scientists from obtaining the essential information they need for their course of study, especially in the context of large-scale genome studies.

In the past few years, the advent of Next-Generation Sequencing (NGS) technologies has given a quick and decisive turning point in response to the ongoing needs of the scientific world, deeply contributing to a better understanding of biological processes. These new methods have influenced the way scientists extract genetic information from biological systems, leading to new fields of study that consider the entire characterization of the studied object [2].

In this context, different areas of science have lead to what is currently called the *–omic* era, which based on use of next-generation sequencing technologies, has dramatically transformed today's biology [3], [4], [5].

The main goal of NGS technologies is producing millions of reads concurrently, in order to decrease the cost and time to obtain biological sequences of interest. For that purpose, different platforms, based on different chemical sequencing approaches, have been developed and constantly upgraded in the last ten years (Table 1), [6], [7], [8].

NGS technologies include a number of methods that widely differ in terms of template preparation, sequencing and data analysis. Every platform presents a unique combination of specific protocols, which determine the type of data produced. The output generated by each platform is different, thus comparing all the available NGS technologies is still a challenge. In fact, even if every NGS Company provides an appraisal related to data quality and cost, there is no general agreement that can establish if a "quality output" from one platform is equivalent to that from another platform [2]. Each platform has strengths and weaknesses.

For instance, if the user is looking for the most throughout per hour, then Ion Torrent PGM would be the best choice [9]. If the goal is to obtain the highest number of sequences in a single experiment, then the best choice would be the Ilumina MiSeq [9]. In case the experiment is based on generating long reads, the Roche 454 is best [8]. All of the instruments use different kinds of PCR (polymerase chain reaction) as preliminary step before the sequencing experiment. This lead to a basic issue associated with the sequencing chemistries that is not possible to stem. Generally two major sources of errors are generated with NGS platforms, which are related to nucleotide substitutions (the instrument reads an incorrect base) and indels due to homopolymers (insertions and deletions of incorrect base), [10].

The main reason why scientific groups decide to adopt one NGS technologies rather to another, basically resides in the cost and the type of data they want to generate (i.e. mostly related to the number and the average length of

sequences and the cost), [11]. A summary of next generation sequencing platforms and their pros and cons is reported in Table 1.

Table 1. Comparison of next-generation sequencing platforms.

Method	Single-molecule real-time sequencing (Pacific Bio)	Ion semiconductor (Ion Torrent sequencing)	Pyrosequencing (454)	Sequencing by synthesis (Illumina)	Sequencing by ligation (SOLiD sequencing)	Chain termination (Sanger sequencing)
Read length	5,000 bp average; maximum read length ~22,000 base	Up to 400 bp	700 bp	50 to 250 bp	50+35 or 50+50 bp	400 to 900 bp
Accuracy	99.999% consensus accuracy; 87% single-read accuracy	98%	99.9%	98%	99.9%	99.9%
Reads per run	50,000 per SMRT cell, or ~400 megabases	up to 80 million	1 million	up to 3 billion	1.2 to 1.4 billion	N/A
Time per run	30 minutes to 2 hours	2 hours	24 hours	1 to 10 days, depending upon sequencer and specified read length	1 to 2 weeks	20 minutes to 3 hours
Cost per 1 million bases (in US$)	$0.75-$1.50	$1	$10	$0.05 to $0.15	$0.13	$2400
Advantages	Longest read length. Fast. Detects 4mC, 5mC, 6mA.	Less expensive equipment. Fast.	Long read size. Fast.	Potential for high sequence yield, depending upon sequencer model and desired application.	Low cost per base.	Long individual reads. Useful for many applications.
Disadvantages	Moderate throughput. Equipment can be very expensive.	Homopolymer errors.	Runs are expensive. Homopolymer errors.	Equipment can be very expensive. Requires high concentrations of DNA.	Slower than other methods. Have issue sequencing palindromic sequence.	More expensive and impractical for larger sequencing projects.

1.2 Next generation sequencing applications

Next generation sequencing methodologies may be applied in a wide range of fields by providing the order of individual nucleotides in DNA or RNA molecules (A, C, G, T, and U) that can be isolated from cells of animals, plants, bacteria, or virtually any other source of genetic information. This has led to increase our scientific knowledge in a quick and effective way, which was unthinkable to obtain before NGS approached the market. For instance, in the molecular biology area, new genomes have been annotated, associating diseases and phenotypes with the possibility to design new drugs for specific biological targets. New studies comparing different organisms and how they evolved have highly advanced our knowledge in evolutionary biology. Biomedical research, ecology, epidemiology, forensic identification, and the biotechnology industry in all its applications have had a remarkable improvement with the use of NGS. Since those approaches have been adopted from different research groups in the entire world, a variety of new fields of biology have been developed, mainly based on what kind of biological molecule is analyzed and sequenced.

For instance, a DNA based approach level will most likely lead to study in the area of:

1. De Novo Sequencing
    - Annotation of genetic code and assembly of novel genomes

2. Whole genome sequencing
    - Discovering the genetic variations in a genome-wide range

3. Amplicon and target region sequencing
    - Finding novel variants or validate candidate variants in a target region

On the other hand a RNA based level will focus on studies in the area of:

1. Transcriptome Sequencing
    - Annotation of the whole transcriptome for the analysis of differential gene expression
    - Discover of novel genes
    - RNA editing analysis (alternative splicing, gene fusion)
    - Discover disease-related functional genes

2. NON coding RNA sequencing
    - Analysis of miRNAs and their regulatory networks
    - Discover disease-specific biomarkers
    - Study of long non coding RNAs

These are just few examples of the effective analysis power offered by NGS technologies. Lately genomics and transcriptomics studies have been the most promising and interesting fields of study in biomedical research, contributing to the annotation of genetic codes of many different organisms (*De Novo* Sequencing), and highlighting the mechanisms that regulate post-transcriptional gene expression (transcriptomics). However, most recently, new different fields of study have been developed, concurrently to the idea of considering our bodies as complex-ecosystems, in which thousands of different species coexist at the same time. The delicate balance that accompanies the simultaneous presence of multiple species in the same habitat

has increased the interest of studying the totality of microorganisms present in a given biological system. This approach is a recent field in biology commonly known as metagenomics [12]. A summary of the main next generation sequencing application is reported in Figure 1, [77].

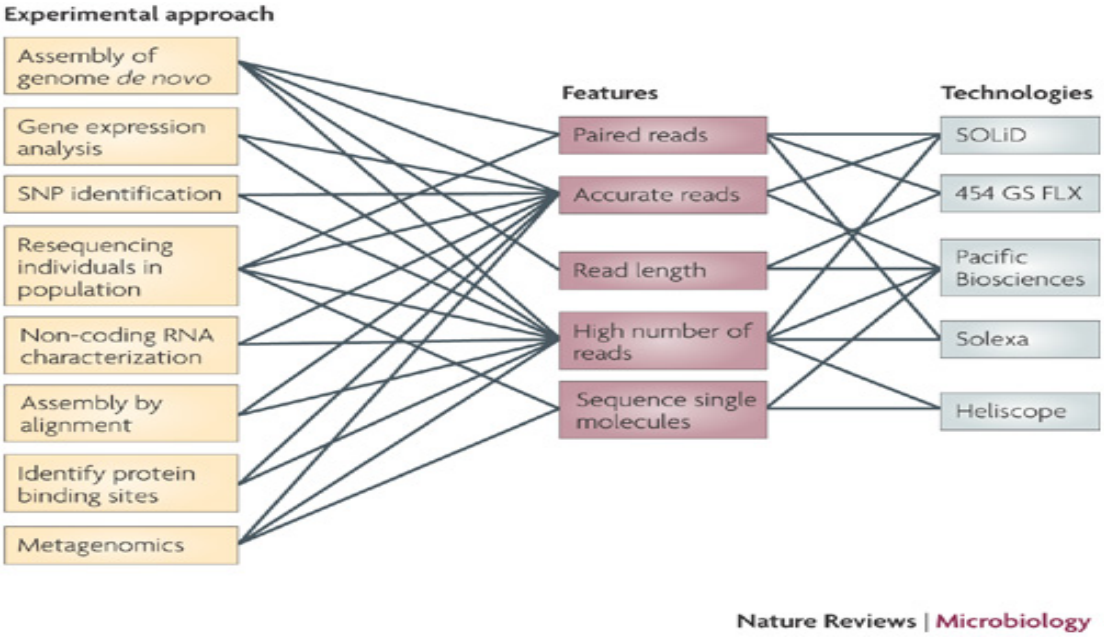

Figure 1. Summary of Next Generation Sequencing applications.

Over the past few years next generation sequencing technologies (NGS) have been applied in a variety of contexts, including whole-genome sequencing, gene expression analysis, SNP discovery, non-coding RNA expression profiling and metagenomics. Those applications have transformed today's biology, allowing scientists to increase knowledge in the area of biomedical research, ecology, epidemiology and forensic identification, with new low-cost and fast sequencing methods. NGS approaches have lead to the development of new field of studies leading to what is currently called the *–omic* era.

(http://en.wikipedia.org/wiki/DNA_sequencing#cite_note-quail2012-3).

1.3 Metagenomics and the human microbiome

Metagenomics is officially defined as "the application of modern genomics techniques to the study of communities of microbial organisms directly in their natural environments, bypassing the need for isolation and lab cultivation of individual species" [13]. Basically, this new area of study related to NGS methodologies, offers the possibility of sequencing the totality of microorganisms in a given environment. The need to develop this new approach relies mainly within the limits of the standard cultivation-based approaches used in microbiology, which have been shown to miss the vast majority of microbial biodiversity [14]. The application of metagenomics allows the deep study of biological systems, and is not limited to the biomedical area, but may be applied to agriculture, ecology and engineering [15].

In this context, a microbiome is defined as the totality of microorganisms, including their genetic elements (genomes), and environmental interactions in a given environment [16], [17]. Is well known that hosted microorganisms living in our bodies play a central role in human health, but they are still several unclear factors, which do not allow to completely highlight the composition and functions of the human microbiome.

The human microbiome is represented by the totality of the microorganisms coexisting in our body, including bacteria, fungi and archeae, and only in the last five years researches are asking questions related to the association of the microbiome alteration and human health [18]. Interestingly, is well known that humans born without any microorganism, but soon after birth a rapid colonization of different body tracts occurs (skin, oral/respiratory tract, genitourinary system and gastrointestinal tract).

An average adult body may contain up to 100 trillion microbial cells, which means there are at least ten times as many bacteria as human cells (~ 10^{14} versus 10^{13}) [19], [20].

The whole human microbiome may constitute the 1-3% of the total body mass [21]. Even if bacteria can be found in all exposed regions of the human body (skin, eyes, nose, mouth) the majority of them is localized along the intestinal tract.

1.4 Role of the human gut microbiome

Although the human genome codes approximately 23,000 genes, this is not sufficient to carry out all the body's biological functions [22], [23]. In fact, only the bacteria of the human gut may encode 3.3 million genes, meaning that a strong interactive influence occurs between human and bacteria gene products [24]. Different studies show the importance of the microbes in the gastrointestinal tract and the interest in this area has rapidly increased since researches first described differences in the composition of the gut microbiome associated with inflammatory bowel disease [25]. The human gut is rapidly colonized after birth and the biodiversity of the microbiome plays a role in the development of gut morphology and physiology. After 2 years of age, the gut microbiome starts to be closer in composition to an adult gut microbiome than to the one of an infant [26]. Thus, changing in microbial colonization of the gastrointestinal tract may represent a major risk factor in the development of food-related autoimmune diseases [27]. The gut microbiome is involved in many essential processes, related to metabolic functions, immunity and development of diseases (Table 2). Those are only some of the basic functions associated with the human gut microbiome. Unfortunately, while we share mostly similarity in the human genetic code, the microbiome seems to be really different from one individual to another.

Many studies tried to draw a picture of the different composition of the human gut microbiome across age, population and geography besides health and disease condition. For instance, is known that common patterns are present in the gut microbiome composition during life, demonstrating that the bacterial diversity is higher in adults compared to children, but in children there are more interpersonal differences [42]. Interestingly, it has been demonstrated how the human gut microbiome differs across different geographical locations. Obviously, the daily dietary is a key factor in influencing the composition and diversity of the microbiome. Consequently, cultural differences may affect the composition of the microbiome because of different dietary habits. Thus, an isolated population of Latin America will most likely have a different microbiome composition to one in Europe [43]. However, genetic factors are less important than dietary factors in determining the microbiome diversification. This hypothesis has been confirmed studying different families, where the microbiome composition presented a similar trend, even when the family members were not directly genetically related (i.e. husband-wife) [42]. Studies in mice have also shown how the composition of the gut microbiome may be influenced.

For example, using germ-free mice scientists have modulated the bacteria diversity in mice guts using different type of diets. Furthermore, other studies have been characterized via microbiome transplantation by colonization of germ-free mice guts with bacteria isolated from human twins discordant for obesity. Surprisingly, mice developed similar symptoms associated to obesity that were present in the affected twins were the microbiome was isolated [44].

Similar studies using germ-free mice have evaluated the importance of the gut flora in response to antibiotic and during inflammation, reshaping the gut microbiome with bacterial transplantation and antibiotic intake [45].

Table 2. Essential processes involving the human gut microbiome.

Metabolic	Immunity	Disease
Positive control on intestinal cell proliferation [28]	Promoting development of gut's mucosal immune system [31], [32], [33]	Prevent tumor formation [31]
Mediate vitamins syntheses [29], [30]	Reduce reaction in allergies and auto-immune disease [34]	Key role in obesity [38], [39]
Absorption of ions [28]	Preventing inflammation [35], [36]	May play a role in mental health [40]
Removal of biochemistry end products and dietary carcinogens [28]	Influence the oral tolerance [37]	Central role in inflammatory bowel disease [25], [32], [33], [41]

CHAPTER 2

THE HUMAN GUT MICROBIOME IN COMMON DISEASES OF THE HUMAN GUT

2.1 Inflammatory bowel disease

Inflammatory bowel disease (IBD) represents a group of chronic inflammatory conditions that occur in the gastrointestinal system. They are considered autoimmune disorders since the body's immune system is involved, targeting the digestive system [58] [59]. IBD is a multifactorial disease that includes interaction of environmental and genetic factors. The rule of the microbiome seems to be crucial and alterations to enteral bacteria can contribute to IBD onset and progression [60][61]. Unfortunately, the genetic factors are still not totally defined, but around 300 genes have been identified as being involved in IBD [62].

Two main forms are really common in IBD affected people: Crohn's disease and ulcerative colitis. Those two disorders basically differ in the location and the nature of inflammation. Another associate disease, even if not classified as IBD, is Celiac disease. In fact, inflammatory bowel disease (IBD) and Celiac disease are both autoimmune disorders that affect digestion and food absorption, and cause similar symptoms but they are not classified in the same group. In fact, while IBD is a term for both Crohn's disease and ulcerative colitis (two diseases that cause adverse autoimmune responses in the digestive tract), Celiac disease instead has a definite cause and effect, which is related to the ingestion of gluten. A person affected by IBD could also suffer from Celiac disease or other kind of gluten sensitivity but they are classified as two distinct diseases with different causes and effects.

It is unclear the cause of the irregular autoimmune response in IBD since it's a complex of genetics and environmental factors. Surely though, IBD is not associated with any food ingestion, as for gluten in celiac Disease.
However, Crohn's disease, ulcerative colitis, other IBDs, and celiac Disease are considered autoimmune disorders affecting the gastro intestinal tract.

2.2 Celiac disease

Celiac disease (CD) is one of the most common autoimmune-disease based disorders, triggered by the ingestion of cereal gluten (a protein find in wheat, rye, and barley) in genetically susceptible individuals. People affected by CD suffer damage to the villi in the *lamina propria* and crypt regions of their intestines when they eat specific food-grain antigens. CD can develop at any age after an individual starts eating gluten and symptoms include abdominal pain, chronic constipation, diarrhea and some time anemia and general fatigue. Other disorders caused by nutrient deficiencies, such as vitamin deficiencies, may develop due to malabsorption [46]. As a result of increased screening in the population more people are daily diagnosed of CD. About 1% of the population in the US and Europe is affected by CD while less people are diagnosed among the African and Asian population due to different genetic background. Furthermore, CD seems to be twice as frequent among female than male [47], [48], [49].
Is well known that a genetic predisposition plays a key role in CD that appears to be polyfactorial, where more genetic factors are involved and more than one is necessary to cause the disease. Two allelic variants have been widely associated with CD: HLA-DQ2 or HLA-DQ8 and 96% of affected people have one of the two HLA-DQ protein types [50], [51]. HLA-DQ belongs to the human leukocyte antigen (MHC-II) system that is an essential element in

the immune system in order to discriminate between *self* and *non-self* cells. DQ2 and DQ8 are associated with the risk of developing CD since the receptors formed by these genes may strongly bind gliadin peptides more than any other antigen-presenting receptor [50]. HLA-DQ2 allele is common and carried by approximately 30% of Caucasian individuals. Nonetheless, HLA-DQ2 or HLA-DQ8 is necessary for CD development but is not sufficient, even if there is an estimated risk of ~36% [47].

Being CD a digestive and autoimmune disorder, it occurs in the small intestine, which is formed with a carpet of small finger-like extroflexions called villi and even smaller projections called microvilli. These structures are of paramount importance for food absorption and to increase intestinal surface so that the body can absorb more essential nutrients (carbohydrates, proteins, and fats), vitamins, and minerals necessary for the daily body diet. When a CD affected subject eats food-containing gluten, after the early stages of digestion, it is metabolized in two smaller components: gliadin and glutanin. The majority of the problem related to CD is attributed to gliadin because this compound is resistant to membrane proteases actions in the intestine, remaining intact after the rest of the gluten is digested. Specifically, gliadin is absorbed in the intestinal epithelium and progresses into the *lamina propria*. At this point the enzyme trasnglutaminase deanimated the gliadin, which changes its form increasing its capability to cause an immune response. This is a crucial point since the new compound, which we know being harmless, is marked as *non-self*, meaning that the immune system marks it as potentially dangerous for the body. Due to this reason, the APC (antigen-presenting cells) bind the gliadin using HLA-DQ2 or DQ8 receptors. The gliadin is then presented to gliadin-reactive CD4 T cells through a T-cell receptor.

The T cells mediate release of cytokines, such as intereleukin-15, responsible for activating the immune system and intraepithelial lymphocytes.

The overexpression of interleukin-15, different cytokine classes and molecules related to the inflammatory cascade causes a progressive destruction of the cells in the intestinal villi. After this process, the T cells and B cells form memory cells, which will trigger the cycle of inflammation and progressive villous atrophy every time a person affected with CD will eat gluten, causing in time a chronic inflammation. Currently, there is no official and definitive cure for CD. The only recognized treatment is following a lifelong gluten free diet, limiting the onset of symptoms [52]. In fact, once gluten is removed from the diet, intestinal inflammation gradually disappeared within a few weeks, although a new exposure to gluten, can lead to a rapid relapse. Even if CD is totally compatible with life (as long as the affected subject follows a gluten free diet) the quality of life itself is dramatically compromised, leading also to social problems suffered by affected people [53]. Particularly, teenagers with CD face different social and school issues related to their condition. This affects newly diagnosed teens, which need to start following a restricted gluten-free diet, but also teens that have been diagnosed earlier in their life since the gluten sensibility is lifelong.

Furthermore, following a gluten-free diet represents problem for the families of affected people since the food is very expansive. In fact, gluten-free substitute foods cost an average of 240% more than their wheat-based counterparts. In fact, gluten-free diet includes fresh fruits and vegetables as well as unprocessed foods, which are more expensive than processed foods.

2.3 Crohn's disease

Crohn syndrome is one of the most common IBD, caused by interactions between multifactorial elements such as environmental, immunological and bacterial factors in genetically susceptible individuals [54], [55].

The anatomy location affected by Crohn's disease may be in any part of the gastrointestinal tract, comprising all the digestive system.

Even though, most frequently it occurs at the ileum level and the beginning of the large intestine. For that reason, symptoms may vary according to the disease location, but generally are similar to CD including: abdominal pain, diarrhea, fever, nausea, vomiting and rectal bleeding [57]. Complications associated with Crohn's syndrome are more common and severe than those associated with CD. In fact they may include obstructions of the intestinal tract and increased risk for colorectal cancer.

As for CD, Cronh's disease may occur at any age, but there is a prevalence of incidence during the adolescence (14-17 years old), although not rare are the cases diagnosed in the range of 50-70 years old [54]. The annual incidence of Crohn's disease ranges from 1 to 10 cases per 100.000 people annually and contrary to CD it is equally distributed among males and females. Interestingly, people who use tobacco products are two times more likely to develop Crohn's disease and healthy siblings of affected people have been reported to have an higher risk to develop Crohn's disease [63], [64]. Similarly to CD, there is no cure for Crohn's disease, and common treatments are only based on controlling symptoms and avoiding relapse. Contrary to CD though, Crohn syndrome is a better understood disease in which clear is the relation that links genetic risk factors with the immune system [56]. In fact, more than thirty genes have been associated with Crohn's disease, especially *NOD2* gene and its variations, because the resulting protein products sense bacterial cell walls.

Crohn's disease pathogenesis seems to be primarily caused by a deregulated proinflammatory response to commensal bacteria and mutations in those related genes might disrupt mucosal defense mechanism. Mucosal defenses are really important mechanisms in order for the body to maintain sterility of the intestinal crypt.

They include barriers like mucus-coated epithelium and the secretions of IgA and defensins that are naturally antibiotics produced by *Paneth cells* that along with *Goblet cells*, enterocytes, and enteroendocrine cells, represent the principal cell types of the epithelium of the small intestine. Being *NOD2* directly involved in the recognition of bacterial peptidoglycan, mutations may affect its function and cause a decrease in defensing production. The progressive decrease and depression of defense mechanisms causes an uncontrolled microbial proliferation with the consequent production of pro-inflammatory molecules (cytokines, interleukins and chemokines) that amplify an abnormal inflammatory response causing the appearance of symptoms [66], [67]. Scientists agree that the inability to control bacteria proliferation in the intestinal walls causes microorganisms to take advantage of the host mucosal layer that are most weakened in affected people, compared to the healthy condition. Also thanks to the use of antibiotics targeting different bacteria strains, a particular resistance has been highlighted in Crohn's disease, leading researches to think that different pathogens are involved [68]. In opposite to CD, Crohn's Disease is not due to an abnormal reaction to specific foods, even though following a specific diet may help reducing symptoms and promote healing. Generally, people affected by Crohn's disease find that soft, bland foods cause less discomfort than spicy or high-fiber foods.

2.4 Influence of the gut microbiome in common human gut diseases

Over the past years scientific research has contributed to highlight our understanding of the role of microorganisms inhabiting human body in health and disease conditions. Our body is a complex ecosystem where distinct populations of organisms coexist, belonging to each of the three domain of life: Archeae, Bacteria, and Eukarya [69]. An alteration in the microbiome composition is commonly known as dysbiosis [78].

Currently, the human gut microbiome and associated dysbiosis is an interesting field of study, describing the role of microorganisms in numerous diseases, including IDB, CD, metabolic disorders, cancer and infections. Different types of interactions occur within our cells, the intestinal mucosa and the hosted species in our body, establishing a delicate host-microbial mutualism. For example, the intestinal epithelium interacts with the gut microbiome providing nutrients in form of mucus, in order to support bacterial metabolism. On the other hand, there are several human genes which products are associated with the development of IBDs since they are part of important pathways. Alterations in the balanced relationship between host and microbiome can lead to an uncontrolled inflammation. Not surprisingly, the incidence of intestinal diseases has rapidly increased over the past few decades, primary due to alterations in microbial environment. For this reason, changing in different aspects of our environment during time, including dietary habits, increasing of vaccinations and antibiotics, together with changes in different aspects of modern lifestyle (living conditions, sedentary life, food processing) have dramatically influenced the microbiome composition. Currently we have only explored approximately the 40% of the gut microbiome that is still uncultured [70]. Is well defined that the composition of a healthy gut microbiome is characterized by four major bacterial phyla: *Firmicutes*, *Bacteroidetes* and less represented *Proteobacteria* and *Actinobacteria* [71], [72]. Globally, different studies have observed dysbiosis in the gut microbiome of IBD patients, mainly regarding a decreased biodiversity, a less presence of *Firmicutes* and an increase of *Gammaproteobacteria* [73], [74], [75], [76]. In opposite, CD is unique among autoimmune diseases since there is a double factor triggering the pathology, represented by both genetic elements and gluten. Different environmental components other than gluten are thought to influence CD development and are still poorly understood.

The interesting fact is that less than 10% of individuals with an increased genetic susceptibility develop CD clinical conditions, and most of them present symptoms even years after their first exposure to gluten. This evidence suggests that other environmental factors could be involved in the pre-autoimmune process, besides gluten. The attention of researchers has been focused particularly on the gluten T-cell specific response. Some assumptions regarding the mimicry of microorganisms proteins and gluten have been made, since our body is continuously exposed to exogenous elements. In this context, some rod-shaped bacillus have been directly associated with the mucosa in CD affected people during the actual disease activity and during the inactivity (for example when patients are under a gluten-free diet) but not in healthy controls [79]. Different studies reported an overall higher incidence of Gram-negative and pro-inflammatory bacteria in duodenal microbiota of CD affected children compared to healthy controls and symptom-free CD patients, suggesting a direct link with the symptomatic presentation of the disease [80]. Several bacteria groups have also been identified to be related with CD, some of those are *Bacteroides*, *Escherichia coli* and *Clostridium leptum*, significantly more abundant in CD patients with active disease than in healthy controls, although these bacterial deviations are normalized in symptom-free CD patients [81]. A similar scenario occurs for the *Staphylococcus* abundance [82]. Less presence of *Lactobacillus-Bifidobacterium* has been reported in CD patients with either active or inactive disease compared with controls. Lastly, in the majority of the available studies, similar bacterial groups have been related to CD both in biopsies and feces [83], [84]. This is an interesting fact that may indicate how the composition of fecal microbiome partly reflects that of the small intestine in CD patients, and could constitute a convenient biological index of this disorder.

However, this hypothesis is somewhat discordant, since other studies have shown that the microbiome composition deeply changes across different sections of the gastrointestinal system [85], [86]. Being that CD and IBD inflammation occur in pretty well known tracts of the gastrointestinal system, prior to the experimental design choosing the anatomy section to analyze, represents a crucial point in the metagenomics studies of the human gut.

CHAPTER 3

BIOINFORMATICS APPROACHES AND TOOLS FOR METAGENOMICS ANALYSIS

3.1 Classification of metagenomics sequencing methods

The increasing of knowledge related to the mechanisms that regulate the development of biological systems, the onset of diseases and the use of new therapies has produced enormous amount of data, which has inevitably led to a more frequent use of computational methods, in order to extract useful biological information for scientific purposes. Currently, there are hundreds of different databases, which annotate genetic codes, variants, non-coding RNAs, proteins and all relevant biological molecules followed by an even more amount of tools, pipelines, and entire open-source projects which allow the analysis of biological data. The study of metagenomics requires the use of different computational methods depending on the experimental design. Historically, the traditional approach was based on the use of small plasmids, known as bacterial artificial chromosomes (BACs), as vectors for DNA cloning and the DNA was sequenced using Sanger method [87]. After the revolution made by next generation sequencing technologies, thousand of microbial genomes have been annotated and whole microbial environments reconstructed including phylogenetic comparisons, profiling and metabolic reconstructions. The pyrosequencing approach has a much higher throughput and a lower error rate per base sequenced, compared to Sanger sequencing.
In fact, the first metagenomics study has been conducted using pyrosequencing technologies on an ancient Mammoth DNA [88].

Since 2006, after the first metagenomics study was published, several studies have been performed and at the state of the art, we have access to more than 10.000 species, annotated in different biological databases [89]. In terms of data collected only the human gut microbiome gene catalog has identified almost 3.3 million of genes for a total of ~600 gigabases of sequences data [90]. The steps from the data collection to the extraction of relevant biological information represent a delicate process and having the right computational power is a challenge, considering the enormous amount of data daily generated. Two main approaches are currently used for the study of metagenomics: large scale shotgun metagenomics sequencing and amplification of small subunit RNA (16S rRNA or SSU sRNA). The first approach is based on the whole DNA sequencing of fragments extracted from microbial population. This method captures the complete genomes of all organisms in a selected environment and allows an accurate phylogenetic inference and a reconstruction of all bacteria genes in the population. The second approach uses 16S rRNA bacteria sequences, which are highly conserved and can be used as a phylogenetic marker for microbial taxonomic classification. The 16S represents more than 80% of total bacteria RNA, and is perfectly suitable for PCR amplification and sequencing. Obviously, there are advantages and disadvantages deriving from the two approaches. For instance, the shotgun sequencing avoids amplification and cloning bias, which are related to PCR protocols [91], [92]. This is a limit related to the PCR primes and the 16S conserved regions. In fact, those regions during long evolutionary periods may undergo some changes and suffer a loss of hybridization to the probe, resulting in underestimation of evolutionary relationships within the population. In this context, has been demonstrated how the V6 bacteria rRNA region amplification may cause overestimation of species richness.

On the other hand, sequencing of the V4 region has been really useful in building phylogenetic trees [93]. Another limit that is present in 16S amplification compared to shotgun sequencing is known as mosaics. Mosaicism is an occurrence due to bacterial horizontal gene transfer. This process, like every significant genomic rearrangement, can not be reported using 16S rRNA region amplification, and it may occur in 16S regions during bacteria genes transfer [94], [95]. Thus, the 16S approach may lead to misidentification, since the identified marker could be a transferred gene. Furthermore, there is no consistent relationship between the 16S rRNA conservation and the total bacteria genome, especially at strain level identification, where two bacteria strains belonging to the same genus may share 98% of 16S rRNA but only 30% of the total genome [96]. On the other hand though, 16S rRNA approach focuses on a small part of the microbial genome, dramatically reducing the sequencing cost. This approach has been particularly effective in monitoring fluctuations in populations. Furthermore, due to the easiness of the techniques prior to the sequencing experiment, this approach is still the most used for large-scale microbiome studies. In fact, multiple samples can be used in the same experiment leading to describe a global taxonomic profile and associated phylogenetic relations within a selected environment.

3.2 Metagenomics shotgun sequencing analysis

Every next generation sequencing analysis starts with a pre-quality filtering procedure, and in this context, metagenomics data need to pass specific pre-filtering steps. Generally, those filters are related to removal of redundancy, low-quality sequences and the total isolation and elimination of any eukaryotic sequence. These steps are really important especially considering metagenomes of human origin.

Filtering and normalization techniques are mandatory for the second step of metagenomics shotgun sequencing analysis that consists in the sequences assembly. Once the data set passes the pre-filtering steps, an assembly procedure occurs. Assembly is the primary goal in shotgun sequencing analysis [97]. A common approach is extracting homologous sequences and assembling them with a comparative assembler or an alignment tool. This is really useful, especially when a fully sequenced genome is available and there are different databases where sequences can be aligned and compared to each other [98], [99]. Not every study has access to appropriate template genomes, especially in the context of non-well studied environment, where non-annotated species are expected to coexist at the same time. In this case, the traditional alignment approaches can not work and instead is generally preferred an overlap-consensus assembly approach. With this method, sequences are clustered in scaffolds (contiguous sequences with known size gaps) with an overlapping approach, where a consensus sequence is built from the totality of the available reads.

This overlap-consensus method has several limits, of which the most commons are: polymorphisms (that are highly present due to different reads from different individuals in the same population) and false overlaps (due to conserved regions shared between species).

To avoid the polymorphisms and the over-laps, generally a single-genome assembly is performed followed by a manual post-processing in order to correct assembly errors. Unfortunately, this method is really time-consuming and depends specifically on the user skills. Currently, new different approaches are being tested, including co-assembly (that uses close related genomes at the same time to correct scaffold errors), even though no promising results have been obtained so far. Once the assembly phase is completed and scaffolds have been obtained, then a gene prediction step occurs.

Depending on the goal of the analysis two distinct ways can be used. The first is used when users want to annotate a pre-assembled genome or multiple genomes that have produced large contigs. In this case it is possible to use different existing tools specifically designed for genome annotation, of which one of the most useful is RAST [100].

With the second method, users can directly annotate an entire community using short contigs or reads that have not yet been assembled. Generally, known genes are first identified and then putative genes functions are assigned. Different algorithms have been developed for gene prediction and they can arrive to have a 95% of accuracy, reducing false negative. Some genes will most likely be lost in the annotation process and the way to avoid this limit is performing a BLAST-based search. Although the last method can annotate the majority of the genes in a microbial community, the huge size of metagenomics data set makes this approach computational expensive and very time consuming.

3.3 Amplicon-based metagenomics analysis

The amplicon-based metagenomics analysis focuses on relatively short reads length (~500 nucleotides). DNA-pyrosequencing applications (454) are really suitable for this purpose, especially considering hypervariable regions within small ribosomal-subunit RNA genes (i.e. 16S rRNA genes). The computational power required for analyzing millions of reads can be a major issue, thus the algorithms developed to analyze and compare amplicon-based metagenomics data have been specifically designed to decrease the amount of power required for a complete analysis.

Different tools are currently available for amplicon-based metagenomics analysis: QIIME [101], MG-RAST [102], MEGAN [103] and Mothur [104].

They basically differ in the way the phylogenetic marker genes are detected, the way the taxonomy is assigned and in the diversity analysis metrics. They also differ in terms of usability, since a tool can be characterized by an easier installation and a user-friendly interaction, while another may require a hard installation process, including third party dependencies and only limited to a terminal user interface. Generally, if is not possible to install a tool natively, it means that it is most likely accessible through an on-line server, offering a graphical-user interface (i.e. MG-RAST). Although people with no any computational skills may use those kinds of tools, they are limited in terms of versatility, throughput and amount of data that can be uploaded. However, in a typical amplicon analysis once the raw data have been produced, different pre-analysis steps are performed. Those are universal steps, based on a pre-filtering phase that every tool can perform. One of the major advantages of amplicon-sequencing approaches relies in the possibility to process multiple samples in the same experiment, globally reducing the cost.

Every sample is associated with a specific barcode sequence, unique to that sample. After the data are generated is possible to identify every sample using bioinformatics methods, in a process called demultiplexing. The first part of the analysis is represented by the quality-filtering step, the primer detection and the assignment of every sequence to a specific sample (demultiplexing). In fact, the majority of the NGS platforms generate sequences that begin with the barcode sequence (unique for every sample), which is followed by a linker primer sequences (i.e. a region of the 16S rRNA) and both are automatically removed during the pre-quality filtering step.

The filtering procedure takes into account at least the quality of the reads (using the Phread quality score, which is a measure of the base-rate error, [105]) and different parameters related to the length of the sequences and the number of errors within the molecules linked with the amplified region (i.e. barcodes, primers).

Different variables are tacking into account and can be set as excluding parameters:

- Phread quality score
- Minimum and the maximum sequences length
- Maximum number of ambiguous bases and homopolymer
- Maximum number of primer mismatch
- Maximum number of errors in barcode

3.3.1 16S rRNAs detection, clustering and identification

After the pre-quality filtering procedure, the resulting dataset can be analyzed using a metagenomics amplicon-based analysis workflow. Currently, the best approach in this area of analysis is represented by the phylogenetic markers detection, performed with an OTU-picking procedure. An Operation taxonomic Unit (OTU) is defined as the taxonomic level of sampling selected by the user to be used in a study, such as individuals, populations, species, genera, or bacterial strains [106]. The definition has been made intentionally vague and could refer to an individual organism, a taxonomic group or a set of sequences evolutionary related that share a set of characters. In the area of amplicon-metagenomics, an OTU is defined as a cluster of reads with a predetermined similarity (usually 97%), which is supposed to correspond to a microorganism species. Obviously, the more the taxonomic classification will be categorized, (i.e. at species level), the more the OTU may produce false negative results. This is understandable if we consider that some species may

share sequences with more than 97% similarity and some cluster may be due to artifacts (generally read errors or chimeras). Nevertheless, different new algorithms have been developed specifically designed for the bacterial 16S rRNA detection, that take into account the possible errors generated during the detection. However, the taxonomy assignment up to the genus level is well characterized by the available methods and may give a sufficient and useful global picture in a metagenomics study. After the genus level (i.e. species), false results are most likely to be reported, and the specificity tends to decrease. This is independent from the tool or the phylogenetic marker detector used, and it is mostly due to the high similarity of the marker gene at the species level. For instance, two organisms belonging to the same genus (but of two different species level), may share the majority of the nucleotides in the marker gene sequences and differing only for a few base pairs.

Those bases are necessary to distinguish one species from another and the detectors may fail to associate the correct taxonomy, due to the high similarity in that region or to the presence of chimera that may generate errors.

Most of the metagenomics analysis tools (i.e. QIIME), provide three high-level protocols for the 16S rRNA detection belonging to the OTU picking procedure: de novo, closed-reference, and open-reference OTU picking:

De novo OTU picking

The de novo OTUs picking procedure is recommended when users need to analyze not common marker genes (where a reference sequences collection is not available) but it can not be use when analyzing non-overlapping regions or huge data set (i.e. >10 millions of reads).

It consists in a process where reads are first clustered against each other, usually at 97% identity, without any reference sequence collection. The final data is a set of OTUs formed by multiple reads belonging to different samples.

Closed-reference OTU picking

In the closed-reference procedure, reads are clustered against a reference sequences collection, discarding from downstream analyses all of those reads that do not hit the reference collection. This procedure is useful when comparing non-overlapping amplicons (i.e. V2 and V4 of the bacterial 16S rRNA) and is the fastest method in terms of speed. Moreover is really useful for very large data set but the novel diversity discovering may be affected, since the reads that do not hit the reference database are discarded. Indeed the choice depends on the environment where the sequences have been extracted (i.e. common and well studied environments such as human gut or not common environments such as soils or deep waters). If the environment has already been well characterized, losing the novel diversity would not represent an issue, since we do not expect any new species. In opposite, the closed-reference procedure may be limited when the study is based on a non-well characterized environment, where unknown are the species and thus losing information may affect the real result.

Open-reference OTUs picking

The open-reference OTUs picking is a procedure where reads are first clustered against a reference sequences collection (same as the closed-reference procedure) but in this case sequences who do not hit the reference database are clustered de novo. This procedure can not be used when comparing non-overlapping amplicons and when users can not provide any reference sequences collection. In this case all the reads are clustered, but the analysis time could be long, especially when using sets with a lot of novel diversity and where sequences are only partially present in the reference database. Although the open-reference procedure speed is faster than the de-novo OTUs picking, it can still take lot of time when analyzing samples with a lot of novel diversity or big data set.

Different studies have compared methods for clustering marker gene sequences into OTUs. In general, a hierarchical clustering method seems to perform better using low dissimilarity thresholds. Furthermore, the sequences abundance plays an important role in the OTU detection. The UCLUST algorithm seems to better handle the limits related to the OTUs picking procedures and currently is the best choice for rRNA clustering and detection [112], [113]. In UCLUST when a cluster is generated, it contains similar sequences based on a similarity threshold, t% identity. Each cluster has a representative sequence (its seed), and all sequences in a cluster are required to have identity ≥t with the seed.

However, there are different available algorithms that can perform the OTU picking procedures with different characteristics, which differ in the way the clustering is performed (Table 3).

Table 3. Clustering methods for the OTU picking procedures.

Clustering Methods	Description
CD-hit [107]	Sequences are clustered using a longest-sequence-first list removal algorithm. Erroneous and chimeric reads are filtered out, combining sequence clustering and statistical simulations.
Blast [108]	Sequences are compared and clustered against a reference database of sequences.
Mothur [109]	For clustering sequences it requires the use of an input file of aligned sequences. The sequences are aligned in a FASTA file to a template sequence alignment.
UCLUST [110]	It creates "seeds" of sequences, which generate clusters based on percent identity; it can take a reference database to use as seeds.
USEARCH [111]	It was developed before UCLUST and works in a similar way but with less general performance.

3.3.2 Taxonomic and phylogenetic assignment

Once the OTU-picking procedure is completed it generates a file, which contains a set of OTUs. Each OTU is formed by a representative set of sequences, without loosing the frequency information. This is really useful to decrease the amount of data to analyze in downstream analysis and in order to reduce the computational power and analysis time. The next step in an amplicon-based metagenomics analysis is represented by the taxonomic assignment. Once all the OTUs are obtained, they need to be associated with specific taxa in order to have a picture of the microbiome composition, based on marker gene amplicons. Assigning taxonomy to a representative set of sequences may be performed with different algorithms and databases, depending on the studied species. Concerning the different methods that, given a set of sequences, attempt to assign the taxonomy of each sequence, the most used are: The Ribosomal Database Project (RDP) classifier [114], BLAST [108], RTAX [115], and Mothur [109]. According on which amplicon region has been used in the experimental design, different databases may be used against the identified groups of OTU, in order to perform the taxonomy assignment. When using common applicon regions, like 16S rRNA sequences, there will be more available databases, previously confirmed with other methodologies and generally more accurate. One of the most cited for the purpose of assigning taxonomy to 16S rRNA sequences is the Green Gene database [116]. Using not common amplicon regions or non 16S rRNA sequences, generally custom databases can be used for assigning taxonomy. Most likely though, those databases have yet to be confirmed and will not have any accurate phylogenetic classification. Having this information is really important for diversity analysis of microbial communities when phylogeny inference is necessary. A summary of the different methods for taxonomic assingment can be found in Table 4.

Table 4. Common methods to assign taxonomy to OTUs.

Taxonomy assigning methods	Description
RDP classifier [114]	This method assigns taxonomy by matching sequence segments (8 nucleotides) against a pre-built database of previously assigned sequence. The quality scores provided by the RDP classifier are confidence values.
BLAST [108]	This method assigns taxonomy by searching input sequences against a blast database of pre-assigned reference sequences. The quality scores assigned by the BLAST taxonomy assigner are e-values.
RTAX [115]	The taxonomy assignment is made by searching input sequences against a FASTA database of pre-assigned reference sequences. Every match within 0.5% of identity is collected. When more than half of the collected matches agree, then the taxonomy assignment is reported.
Mothur [109]	Similar to the RDP Classifier, this method requires a set of training sequences and associated id-to-taxonomy assignments.

Inferring a phylogenetic tree relating the sequences

Simultaneously to the taxonomic assignment, in an amplicon-based metagenomics analysis pipeline, it is common to generate a phylogenetic tree to correlate the sequences. This step is necessary since the tree is used with phylogenetic tools involved in downstream analysis (i.e. diversity analysis). Generally the alignment can be performed *de novo* or assigning sequences to an existing template alignment, on the basis of which the tree will be generated. The difference between *de novo* alignment and the pre-built alignment resides basically in the quantity of sequences involved in the study. For instance, for small studies (with less than 1.000 sequences) a *de novo* alignment can be performed using specific tools. One of the most useful for this purpose is MUSCLE [117]. For studies involving more than 1.000 sequences, is preferable to use a pre-built alignment methods, since the *de novo* aligners would be too slow. Thus, in large studies, other tools are generally involved. The most used aligner tool using a pre-built template alignment is PyNast [118]. Template alignments for common amplicon sequences are widely available (i.e. 16S rRNA). On the other hand, if non-common amplicon are involved in the study, a template alignment is not available (i.e. 18S rRNA and ITS) and the *de novo* option does not lead to accurate results. This is due to both the huge time required to perform a *de novo* alignment and the computational power involved. Furthermore, not so many studies involving non 16S amplicon regions have been performed in the metagenomics area, therefore having a phylogenetic tree based on input sequences set is still a challenge. This is a limit when diversity community analysis involves phylogenetic relationship. The final output will be a phylogenetic tree. Metagenomics tools generally generate trees in Newick tree format and they can be visualized with several tree visualization softwares [143].

3.3.3 Basic input and expected analysis output

The Meta data file: A collection of information used as analysis input

In a large-scale metagenomics analysis, it is mandatory to have all the information related to each sample in order to address biological questions. In this context, regardless the metagenomics tools, the user is supposed to provide a meta-data mapping file. This file should contain all the information related to the samples necessary to perform the data analysis. The file should report a list of variables, provided by the user, associating every sample or group of samples with a category. For instance they could be information related to the patients (i.e. age, sex, disease status, geographical origin, dietary habit, clinical background) or related to the experiment (i.e. library kit used, NGS platform, amplicon region). The more variables are reported, the more will be easy to address biological question during downstream analysis. At least the meta data file should contain the name of each sample, the barcode sequence used for each sample (useful for multiplexing and to associate every single sequence to the sample it belongs) and the linker/primer sequence used to amplify the sample.

Output reporting the taxonomic composition: The Biological Observation Matrix

During an amplicon metagenomics analysis, once the taxonomic assignment is completed, it is necessary to generate a readable output reporting all the taxa associated with the samples. The output should take into account, the different taxonomic relationship within every sample and the frequency of which a taxon occurs in a sample. The best way to represent this type of data is a matrix that relates samples and taxa reporting the respective frequencies.

At the state of the art, all the metagenomics analysis tools are going in a common direction in terms of output generated in their pipelines. The current universal recognized standard is the Biological Observation Matrix (BIOM) format [119]. The BIOM format is specifically designed for representing biological sample by observation contingency tables. It is a recognized standard for the Earth Microbiome Project (*http://www.earthmicrobiome.org/)* and a candidate for the Genomic Standards Consortium (*http://gensc.org/).* It was first developed to facilitate the handling of large-scale comparative *–omic* data, allowing the storage of contingency table data and sample/observation metadata in a single file. Secondly, it was developed to facilitate the use of taxonomic tables between the available tools in order to create a common standard format. The primary use of the *biome* table format is to represent OTU tables in a metagenomics analysis. In this case, the observations are OTUs and the matrix contains counts corresponding to the number of times each OTU is observed in each sample. The function of the *biom* format is not limited to the storage of taxonomic composition in a metagenomics study. In fact, the table can be used in many different contexts. For example it can be uploaded in external tools in order to generate plots representing the taxonomic profile; it can be used within different algorithms to perform mathematical operations (i.e. filtering or summary). Furthermore, it can be used to compute statistical analysis and it is a standard format necessary to perform community diversity analysis. Currently, all the available metagenomics analysis tools recognize the *biom* format and even new algorithms, are integrating the *biom* table as a standard output format.

3.3.4 Diversity analysis

One main goal in metagenomics analysis is describing the microbial diversity within a study. In ecology, different type of species within a community and their abundance at a specific scale represent the biological diversity. Two common terms for measuring biodiversity have been described: alpha and beta diversity [120].

Alpha diversity

Alpha diversity, computes the diversity within a particular area or ecosystem, and is represented by the number of species (species richness) in a biological system. Basically the alpha diversity tries to answer different questions related to the community richness, such as: How different is the composition in every sample presents in a dataset? How many different species are present in every single sample? How does the number of sequences in every sample influence the species richness? The first definition of alpha diversity was introduced in 1972 by Whittaker as the species richness of a place [120]. However, the practical development of this concept has been redefined tacking into account the structure of the community. The most common expression links both the number of species and the proportion in which each species is represented in the community. Basically, if in a community the number of different species is high and their abundances are similar, then the alpha diversity has a high index.

However, before computing alpha diversity, it is necessary to take into account the number of sequences generated for every sample. In fact, we can expect in a sequencing experiment that if more sequences are generated, then more species will be identified. This may represent a limit, considering that we can have X reads from one sample and 1/5 X reads from another sample.

In fact, we could expect to find more species for that sample if we sequence 5 times more. In order to avoid this problem, before the community diversity is computed, generally a rarefaction step is performed. A rarefaction is basically a random collection of sequences taken from a sample, with a specified depth (related to the number of sequences). For instance, a rarefaction with a depth of 100 reads per sample would be a simulation of what the sequencing result would be if the sequencing experiment generated exactly 100 reads for each sample. Usually, many rarefactions at multiple depths and repeated many times at each depth are performed. This is a common way to normalize the sequences/samples generated after a NGS experiment. The most important factor prior to the rarefaction procedure is to choose the rarefaction depth based on the total sequences per sample.

Generally, the rarefaction depth is the number corresponding to the minimum number of sequences belonging to a sample within the dataset, if and when that number is close to the average of the sequences in every sample. Obviously we need to carefully consider the overall data generated and the number of samples in the study. For instance, we can imagine our data having an average of 10.000 sequences/sample. In case some of the samples (no more than the 5% of the total) have less than 1.000 sequences each, we would choose a high number of rarefaction depth (~10.000), excluding those samples. On the other hand, if more than the 5% of our samples has a sensible different number of sequences compared to the average, we would asses to 1.000 the rarefaction depth in order to not lose the majority of the samples. The goal is trying to find the best rarefaction depth parameter, considering the average of the sequences per sample but also the total number of samples in the dataset. This is important for not excluding samples, but at the same time for not compromising the statistical power.

The rarefaction is computed on the OTU *biom* table generated after the taxonomic assignment, and the results will be multiple OTU tables.

The alpha diversity is computed on every rarefied OTU table and then collapsed in a single table. The final output will be a normalized OTU table reporting the average of the alpha diversity. Several indices have been developed in order to compute the alpha diversity and the choice of a different index depends on the type of data. Every method computes the species richness of every sample considering different parameters:

- The number of observed species per number of sequences/sample
- The presence of singleton (species with only one sequences) or doubletons (species with exact two sequences)
- Phylogenetic distance
- Species evenness (how close in numbers each species in an environment are)

A summary of the main alpha diversity indices is reported in Table 5.

Table 5. Summary of the main Alpha Diversity Indices.

Alpha diversity Index	Function
Observed Species [121]	Is one of the most used alpha diversity index. It compares the number of identified species with the number of sequences in every sample. It takes into account the number of unique OTUs found in every sample.
Chao1 [122]	It computes the species richness by using the number of rare species that are found in a sample, as a way of calculating how likely is the presence of undiscovered species. Basically, it compares the total number of the species found, the number of singleton (species with only a single occurrence in a sample) and the number of doubletons (species with two occurrences).
Shannon Index [123]	Shannon's index accounts for both abundance and evenness of the species (how close are in numbers each species in an environment). It provides estimates of the effective number of species present by including or ignoring the relatively rare species.
Phylogenetic Diversity [124]	The phylogenetic diversity (PD) measures both species abundance and phylogenetic distances. It requires a phylogenetic tree together with the OTU table.

Beta diversity

The beta diversity (diversity between samples) is a term used in ecology for the comparison of samples to each other. It tries to answer basic questions that correlate the samples: How samples from environment A differ to the samples from environment B? How different (in terms of species) are my samples compared to each other? Do samples within the same environment cluster together?

Opposite to alpha diversity, which calculates a value for each sample, the beta diversity computes distances between pairs of samples. The output is a matrix of the distances of all samples compared to all other samples. As for the alpha diversity, sequencing depth can influence beta diversity analysis. To avoid this limit a rarefaction step (as described for alpha diversity) is generally computed prior to beta diversity analysis in order to standardize the data obtained from samples with different sequencing counts. The matrix reporting the distances between every pair of the community samples will reflect the dissimilarity between the samples. There are different metrics to produce the distance matrix. The input on which the matrix is computed is the OTU table reporting the number of sequences observed in each OTU and for each sample. Beta diversity methods are phylogenetic and non-phylogenetic based. Generally at least one phylogenetic metric is supposed to be considered in the analysis, since the result can be vastly more useful to address biological questions [125]. The most common phylogenetic metric for beta diversity analysis is Unifrac [126]. Unifrac is an online tool that allows the microbial communities comparison using phylogenetic information. The input it requires is an OTU table, generated previously in the analysis workflow and a phylogenetic tree containing sequences derived at least from two different comparisons (i.e. different environments or different disease status).

The Unifrac tool helps to determine if our dataset has significant different communities and if those differences are remarkable associated with a particular lineage in the phylogenetic tree. It clusters together samples to assess if some variable (provided by the user in the meta data file) influences the observed clusters. Furthermore, it is also used as a standard to highlight differences between environments that are geographically distant.

Unifrac allows two main different measures: weighted and unweight.

The weighted Unifrac is a quantitative measures ideally suited to reveal community differences that are due to changes in relative taxon abundance. It takes into account the number of sequences and it is useful when some taxa decrease or increase in the environment do to external changes (i.e. a limited nutrient).

The unweighted Unifrac is instead a qualitative measure that does not consider the number of sequences but only the presence/absence of a taxon. This is useful to highlight what species are present in a particular environment (i.e. extreme temperatures, PH, and pressure) and also to avoid the abundance of a particular taxon to obscure other patterns of variation. Usually both the weighted and unweighted Unifrac metrics should be used [127]. When a phylogenetic tree is not available, (i.e. when using non-common amplicon regions), other beta diversity metrics can be used that do not consider phylogenetic relationship.

Non-phylogenetic metrics are less informative than phylogenetic metrics, but are useful to cluster samples within an environment based on the sequence similarity. In this context, one of the most used non-phylogenetic metric is Bray-Curtis dissimilarity [128]. It describes the dissimilarity between the structures of two communities based on the abundance of the sequences.

The Bray-Curtis index is computed for every pair of samples and the computation involves summing the absolute differences between the counts and dividing this by the sum of the abundances in the two samples.
To illustrate its function, we can consider the count of OTUs for two samples, A and B:

	OTU1	**OTU2**	**OTU3**	**OTU4**	**OTU5**	**SUM**
Sample(A)	11	0	7	8	0	26
Sample(B)	24	37	5	18	1	85

The Bray-Curtis dissimilarity index (b) for the pair of samples A, B will be:

$$b_{(sA,B)} = \frac{|11-24| + |0-37| + |7-5| + |8-18| + |0-1|}{26+58} = 0.568$$

This measure is represented on values between 0 (identical samples) and 1 (samples completely different). A matrix is built reporting all the distance values between every pair of samples in the dataset. The beta diversity metrics will generate distance matrices with different methods according to the metric used. The matrices are the basis for downstream analysis, allowing the clustering and visualization of the sample distances (principal coordinate analysis, hierarchical clustering, and distance histograms). A summary of the most common beta diversity metrics is reported in Table 6.
Since the data obtained after the beta diversity analysis are represented by multidimensional data, it can be hard to find patterns between samples. For the purpose of the beta diversity analysis visualization, different methods are used to identify patterns and to highlight similarities and differences.
A standard method to graphically represent the data is Principal Coordinate Analysis (PCoA). It is a technique that helps to extract and visualize

informative components of variation from complex, multidimensional data. Basically the goal of PCoA is to map samples from a distance matrix in a set of orthogonal axes. The principal coordinates can be plotted in two or three dimensions. This is useful to have an intuitive visualization of the data structure in order to find differences between the samples, or looking for similarities by sample category (a variable in the meta data file).

PCoA uses a method, known as multidimensional scaling, that computes a linear transformation of the variables into a lower dimensional space, retaining the maximal amount of information.

Each of the samples will be visualized in a graph, explaining the variability between every sample and within a category. Multiple graphs are generally generated as output, in order to find possible associations (express by clusters) that can be useful to address biological questions. The more variables have been included as input in the mapping file, the more plots will be generated. Every plot will report a cluster of samples within a specific category. In this way, it is intuitive to identify if samples from a category (i.e. affected) cluster together, compared to samples belonging to another category (i.e. healthy). The basic idea is that the more the sample are close in a cluster, the more their microbiome is similar.

Table 6. Summary of the main Beta Diversity Metrics.

Beta diversity Metrics	Function
Unweighted Unifrac	A qualitative phylogenetic metric that considers the presence/absence of a taxon in a phylogenetic tree. Useful in extreme environment studies to avoid the abundance of particular species to obscure other less present species.
Weighted Unifrac	A quantitative phylogenetic metric ideally suited to reveal community differences that are due to changes in relative taxon abundance. Useful when the abundance of some species changes in the environment due to external changes (PH, temperature, pressure).
Bray-Curtis Dissimilarity	A non-phylogenetic metric that describes the dissimilarity between the structure of two communities based on the abundance of the sequences and their similarity between the samples. Useful when a phylogenetic tree is not available due to non-common amplicon regions (i.e. ITS).

3.3.5 Statistical analysis

Sequencing amplicon shotgun regions from microbial communities with NGS technologies generates millions of reads. Because of the huge amount of data and the associated information, different statistical methods are required for the data evaluation. Numerous arithmetic and statistical models are available to assess the composition and diversity of microbial communities. The first statistical analysis is computed using the OTU table. Being this type of file a biological matrix (reporting samples and associated taxa with every single frequency), it is possible to use various approaches to identify statistical significance within the samples and by different categories. For that reason, the meta data file is necessary as initial input. In fact, it contains information that links the variables by different categories. We could be interested in the statistical evaluation of certain identified species, considering different variables. For example, it is possible to identify taxa that are abundant in a specific subcategory of the samples compared to another (i.e. affected patients versus healthy controls, female versus male or young versus old).

To determine the statistical significance of each taxon assigned, the elective tests is the Analysis of Variance (ANOVA) using a Bonferroni correction [129], [131]. ANOVA can find OTU differentially represented across experimental variables, computing a p-value for each taxa and then adjusting the p-value with the Bonferroni correction to avoid the problem of multiple comparisons.

Concerning the statistical evaluation of the diversity analysis matrices, the majority of the currently available comparison techniques are based on the ANOVA family of statistical methods. These methods determine if the grouping of samples by a given category is statistically significant.

The statistical tests used in diversity analysis comparison, are nonparametric and they use permutations of the data to determine the p-value and the statistical significance.

The methods universally accepted for the statistical analysis of microbial community data have been evaluated empirically, using simulated data sets to verify their reliability on microbial community data. Although the methods have been tested, they can present some limits. Due to their nonparametric data, the methods do not assume any normality of the data but instead they assume equal variance of every group of samples. Due to the equal variance assumption they can suffer from low specificity (i.e. detecting significance of some group even when it is not expected). Furthermore, it has been evaluated that the more the number of samples in the study increases, the more increases the possibility to find significant p-values. The more used tests are the Analysis of Similarity (ANOSIM) and the permutational multivariate analysis of variance (ADONIS) [132], [133], [134]. The two tests are similar and they are generally used independently on the same data set in order to see if they agree.

ANOSIM is a method that tests whether two or more groups of samples are significantly different. This is useful when we want to find significant differences that associate samples in groups (based on the variables specified in the meta data file). Since ANOSIM is a nonparametric test, statistical significance is determined through permutations. The test will generate both a R-value and a p-value.

The R-value will be in the range of +1 and 0. This means that the more the value is close to 1 the more there is a strong dissimilarity between the groups. The R-value together with the p-value at an alpha of 0.05 will help indicating if the grouping of samples by a specific category is statistically significant.

The ADONIS test is also a nonparametric statistical method that computes the statistic on a distance matrix file (generated with beta diversity metrics). In this case too, a meta data file is required as input, specifying a category to determine sample grouping. The effect size, expressed as an R2 value, will be in form of percentage of variation explained by the subcategory indicated and extracted from the meta data file. A p-value is also computed determining the statistical significance. The R2 is created identifying the relevant centroids in the data and then calculating the squared deviations from these points. The p-value instead is computed using F-tests on sequential sums of squares from permutations of the data. Thus, specifying a category in the meta data file, the result will be an R2 = X%, indicating that approximately X% of the variation in distances is explained by that grouping. The p-value will define if the grouping of samples is statistical significant.

A complete amplicon-based metagenomics analysis consists of different steps from the raw data to the statistical evaluation. Every single step can be performed with the available metagenomics analysis tools and the pipelines can be modified according with the user needs and the type of experiment. It is important to have a clear idea of what kind of workflow will be used prior to the raw data upload. In fact, even just changing a single parameter in the workflow can dramatically influence the final result. In Chapter 3, I described in details what are the current methods to perform a complete amplicon-based metagenomics analysis.

Regardless the different algorithm belonging to every single analysis step, a complete workflow summary of a general amplicon-based metagenomics analysis is reported in Figure 2.

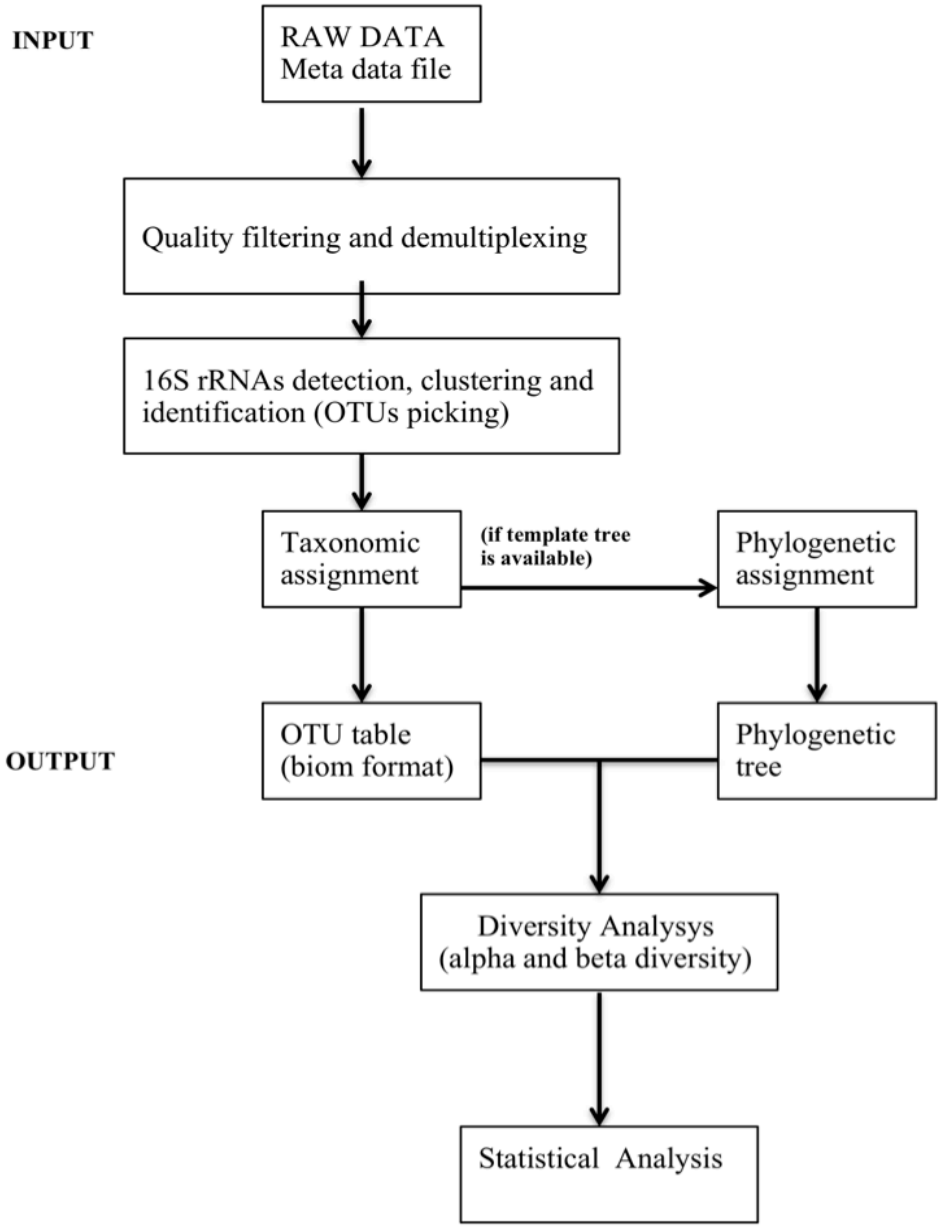

Figure 2. Amplicon metagenomics analysis flowchart.

Essential bioinformatics steps performed during an amplicon-based metagenomics analysis. From the raw input data, the sequences are quality filtered and demultiplexed. From the obtained output, the OTUs are picked and the taxonomy is assigned. The final output is a biological matrix table (*.biom*). The *.biom* table is used as input in order to obtain phylogenetic assignment, diversity analysis and statistical evaluation.

CHAPTER 4

AIM OF THE PROJECT AND MOTIVATION

Based on the evidence, that the gut microbiome plays the same role in common bowel inflammatory diseases and celiac disease, I present two distinct studies, conducted using an amplicon-based metagenomics approach. The first study investigates the microbiome composition in Crohn's disease. As mentioned in Chapter 2.3, its pathogenesis is not well known but recent evidence directly associates the gut microbiome with the development of Crohn's disease. It is also well known that nutritional therapy is effective in children affected by Crohn's disease by inducing remission of mucosal inflammatory reactions with an overall beneficial effect on the child's nutritional status and growth [135]. Recent studies have suggested that nutritional therapy may also modify the fecal microflora in affected children [136].

The aim of the first study was to characterize the gut mucosal microbiome in a child affected by Crohn's disease at diagnosis and after nutritional therapy. For comparison purposes, the analysis was performed also on the microbiome of a healthy child of the same age and sex as the affected patient. The microbiome profile was characterized by 16S rRNA sequencing using a high throughput sequencing approach. In celiac disease (CD) (described in Chapter 2.2) genetic factors are not sufficient to explain the onset of the disease. Other factors such as environmental factors, the innate and adaptive immune system and the intestinal microbiota may play a role in the different manifestations of CD. The role of intestinal microbiota in CD is still unknown, therefore I have analyzed the duodenal microbiota composition of adult patients with CD and matched healthy controls.

Since dietary intake of gluten, seems to be the main CD triggering factor, in this study I have included a group of patients affected by CD adhering to a gluten-free diet (GFD). The goal is to highlight differences in the microbiome composition of the affected patients compared with healthy controls, and also to understand if diet can directly influence the microbiome composition.

The analysis was performed on duodenal biopsy samples of 15 patients with CD, 10 control subjects and 6 CD patients at GFD undergoing CD follow-up. Since no study has characterized the ileum microbiome in CD, and since the microorganisms that may be associated with symptoms are unknown, the study was designed to elucidate both the bacterial and fungal composition.

The microbiome profile was characterized using 16S and ITS (bacteria and fungi) ribosomal RNA (rRNA) gene sequencing together with a high throughput sequencing approach to evaluate if imbalances in the composition of gut microbiota may be related to CD presentation.

CHAPTER 5

MATERIALS AND METHODS

5.1 Patients and sampling collection

In the Crohn's study a 14-year-old boy was enrolled with pediatric active Crohn's disease. After colonoscopy, he started a nutritional therapy. The diet is restricted to a daily powder constituted by proteins, antioxidants, and anti-inflammatory fats for a period of 8 weeks. After this time, a clinical re-evaluation revealed disease remission. The samples were obtained from endoscopic ileum mucosal at diagnosis and after nutritional therapy. Another sample was obtained from the ileum tissue of a non-Crohn's disease 15-year-old boy.

For the celiac disease study, thirty-one unrelated Caucasian individuals were recruited, in a one-year period. Several exclusion criteria for the enrolment have been adopted in order to not alter the gut microbiome composition. The criteria were related to any known food intolerance, IgA deficiency, therapies with antibiotics, antiviral or corticosteroids or assumption of probiotics in the 2 months before the sampling time.

The samples are formed by ileum endoscopy of:

I) 15 subjects (87% females, mean age/range 34/20-51 years, with the exception of a 14-year-old female) with active CD; II) 10 individuals (80% females, mean age/range 33/20-52 years) as clinical controls; III) 6 subjects (83% females, mean age/range 38/25-53 years) with non active CD following a Gluten Free Diet (GFD) from at least 2 years before sampling time.

A summary of the subjects enrolled for both Crohn and celiac disease study is reported in Table 7.

Table 7. Subjects enrolled in Crohn and celiac disease study.

Celiac Disease	Controls	CD-Patients	GFD-Patients
N. of subjects	10	15	6
Age (Mean/range)	33/ 20-52	34/ 20-51[§]	38/ 25-53
Sex (Female/Male)	8F/2M	13F/2M	5F/1M

Crohn's Disease	Controls	Patients*
N. of subjects	1	1
Age (Mean/range)	15	14
Sex (Female/Male)	M	M

*Same subject sampled again after nutritional therapy

5.2 16S and ITS rRNAs amplification and sequencing

For both the studies I used the same amplification approach and sequencing technology. In the Cronh's disease study only the 16S rRNA analysis has been performed, while for the celiac Disease study both 16S and ITS rRNA. Specifically, the total DNA was extracted from duodenal biopsies (3 mg/sample). To assess the quality of the genomic DNA extracted, a gel electrophoresis was performed excluding any RNA contamination or degradation.

The genomic DNA (gDNA), whose quality was assessed by gel electrophoresis, resulted to be free of RNA contamination and degradation. An aliquot of the duodenal DNA was used for PCR amplification and sequencing of bacterial 16S (for both celiac and Cronh's disease studies) and fungal ITS rRNA genes (only for the celiac study). To deeply investigate the bacterial composition of duodenal samples, a 548 bp amplicon, spanning from V4 to V6 variable regions of the 16S rRNA gene, was amplified using specific primer [138]. ITS rRNA amplicons were obtained as previously described [139].

The primers were checked on RDP database reporting ~175.000 match for 16S and ~10.000 match for ITS. After visualization by gel electrophoresis, each PCR products was individually purified, assessed for quality and quantified. Equimolar amounts of each amplicon were pooled together to obtain multiple amplicon libraries (1 library/subject). Every sample was associated with a unique MID identifier and the pool library was loaded in a 454 FLX+ Titanium (Roche) to generate thousands of sequences for further analysis.

5.3 Bioinformatics analysis

All the dataset was analyzed using the QIIME package v. 1.7.0 [101]. It stands for Quantitative Insights Into Microbial Ecology and is an open source software package for comparison and analysis of microbial communities based on high-throughput amplicon sequencing data. A based-amplicon metagenomics analysis was performed, following a specific workflow (described in Chapter 3.2).

5.3.1 Quality filtering, primers detection and demultiplexing

The preliminary analysis steps included quality filtering, primers detections and demultiplexing in order to label every sample by a unique nucleotide barcode identifier. Those steps provide the use of a script integrated in QIIME (*split_library.py*) with the following parameters used for both the studies:

Min average quality Phread score allowed in reads = 25
Minimum sequence length, in nucleotides = 200
Maximum sequence length, in nucleotides = 1000
Maximum number of ambiguous bases = 6
Maximum length of homopolymer run = 6
Maximum number of primer mismatches = 0
Maximum number of errors in barcode = 1.5

The previous steps were performed for 16S bacteria and ITS fungi dataset in both the studies.

5.3.2 Pick Operational Taxonomic Units (OTUs) and pick a representative sequence from each OTU

The quality-filtered sequences were submitted in the QIIME OTUs picking pipeline, using an open-reference OTUs picking approach (described in Chapter 3.2.1). The OTU picking step assigns similar sequences to operational taxonomic units by clustering sequences based on a defined similarity threshold. Sequences that are similar at or above the threshold level are taken to represent the presence of a taxonomic unit in the sequence collection. Different clustering methods have been implemented in QIIME. For both celiac and the Crohn's disease dataset, UCLUST algorithm was chosen for clustering the sequences and for the OUTs picking procedure [110]. UCLUST creates "seeds" of sequences that generate clusters based on percent identity. In this case, a 97% identity was set, enabling reverse strands matching for the OTU picking. The UCLUST algorithm was chosen since it is the most suitable for this type of data [112]. The output from UCLUST was summarized in order to obtain a representative sequence of each OTU without loosing the relative expression level information.

5.3.3 Assigning taxonomic identity to OTU using a reference database

After the OTU picking procedure, a representative list of OTUs was obtained. The representative OTUs were associated to a 16S rRNA and ITS sequences database to identify the taxonomic composition. This step is crucial to provide the correct microbial lineages belonging to each sample. QIIME by default uses different taxonomy assigner. In this case the RDP classifier v 2.2 was chosen because of its flexibility, accuracy and moderate computational power demanding [114].

The 16S OTUs were classified using the green genes database v. 2012, while the ITS OTUs were classified with the UNITE ITS database v. 2012 with a minimum of assign taxonomy confidence of 80% [141], [142]. Using the taxonomic assignment combined with a user built meta data file (containing metadata information such as age, sex and disease status) QIIME assembles a readable matrix of OTU abundance for each sample with meaningful taxonomic identifiers for each OTU. The OTU matrix table is in *.biom* format [119]. The *.biom* format is based on JSON (JavaScript Object Notation). JSON is a widely supported format with native parsers available within many programming languages. More info on *.biom* format can be found here: http://biomformat.org/documentation/format_versions/biom-1.0.html

5.3.4 Aligning OTU sequences, filtering the alignment and building a phylogenetic tree

Alignment of the sequences and phylogeny inference is necessary for subsequently analysis with phylogenetic tools. To perform this kind of analysis a template alignment is necessary to associate microbial sequences within a phylogenetic tree. At the state of the art, for ITS sequences a template alignment is not yet available, thus building a phylogenetic tree for ITS fungi sequences is still a challenge. Currently, different research groups are collaborating to provide an accurate answer. For this reason, the template alignment analysis step was performed only on 16S rRNA sequences. QIIME can perform alignment of the sequences through assignment to an existing template alignment using PyNast [118]. PyNast is a python implementation of the NAST alignment algorithm [144]. The NAST algorithm works aligning every sequence to the best-matching sequence in a pre-aligned database of sequences (the “template” sequence).

PyNast does not allow introducing new gap characters into the template database, so the algorithm introduces local mis-alignments to preserve the existing template sequence. In this case too, a minimum quality thresholds is the main requirements for matching between a candidate sequence and a template sequence.

Default parameters were chosen, setting the minimum sequence length to 150 nucleotides and the minimum percent identity to 75%. Even if pyrosequencing reads are long (~400 bp) in order to be compatible with the NAST algorithm, PyNast by default sets the minimum sequence length to 150 nucleotides.

The sequences were aligned against the Greengenes template alignment [141]. Once the alignment output file is generated, a filtering procedure is highly recommended in order to remove columns comprised of only gaps, and locations known to be excessively variable. This is essential to remove positions which are formed by gaps and that could negatively influence the phylogenetic inference. Finally, the filtered alignment file produced, is used to build a phylogenetic tree in the Newick format using a tree-building program [143].

After these initial steps in the workflow analysis, several files are generated. Particularly the workflow generated OTU tables in *.biom* format (respectively for bacteria and fungi) that contain taxonomic information as well as metadata information related to every sample, for both celiac and Crohn's disease study.

5.3.5 Diversity analysis: Alpha and Beta diversity

Once the OTU *.biom* tables and the phylogenetic tree were obtained, they were combined to compute diversity analysis.

For both alpha and beta diversity, sequencing depth can influence beta diversity analysis. The sequence depth is only related to the results (number of sequences/sample) obtained after the next generation sequencing experiment. To avoid this limit a rarefaction step was computed prior to diversity analysis in order to standardize the data obtained from samples with different sequencing counts. Many rarefactions at multiple depths and repeated many times at each depth were performed. This is a common way to normalize the sequences/samples generated after the next generation sequencing experiment. Rarefaction curves and diversity indexes were calculated using default parameters at a sequence evenness related to the average of sequences. The result is represented by multiple-OTU tables at different rarefactions depth. The alpha and beta diversity are computed on all the multiple tables and finally the result is collapsed into a single merged table.

Specifically, for the alpha diversity analysis I used the (I) Observed Species (22) metric, which is one of the most used alpha diversity index; the (II) Shannon Diversity Index [23] which accounts for both abundance and evenness of the species (how close are in numbers each species in an environment). It provides estimates of the effective number of species present by including or ignoring the relatively rare species. The (III) Chao1 [24] richness estimator, that computes the species richness by using the number of rare species that are found in a sample, as a way of calculating how likely is the presence of undiscovered species. Basically it compares the total number of the species found, the number of singleton (species with only a single occurrence in a sample) and the number of doubletons (species with two occurrences).

For the Beta Diversity I used both phylogenetic and non-phylogenetic metrics. In fact, in order to compute beta diversity based on phylogenetic distances, an input phylogenetic tree is required. Since for ITS fungi sequences a tree is not still available, the phylogenetic beta diversity was computed only on the 16S rRNA data while for the ITS a non-phylogenetic beta diversity index was used. Generally, at least one phylogenetic metric is supposed to be considered since the results can be vastly more useful to address biological questions [25]. Specifically, the goal is to generate a matrix reporting the distances between every pair of the community sample that will reflect the dissimilarity between the samples.

The most common phylogenetic metric for beta diversity analysis is Unifrac [26]. Unifrac is an online tool that allows the microbial communities comparison using phylogenetic information. The input it requires is an OTU table, (generated previously in the analysis workflow) and a phylogenetic tree containing sequences derived at least from two different comparisons (i.e. different environments or different disease status). The Unifrac tool helps to determine if our dataset have significantly different communities and if those differences are remarkable associated with a particular lineage in the phylogenetic tree.

The most common non-phylogenetic metric (used here for computing the beta diversity of the ITS CD data set) is Bray-Curtis [27], which describes the dissimilarity between the structures of two communities based on the abundance of the sequences. Both the metrics produced a matrix.

The data obtained after a beta diversity analysis are represented by multidimensional data collected in a matrix and it might be hard to find patterns between samples. For the purpose of the beta diversity analysis visualization, different methods are used to identify patterns and to highlight similarities and differences.

A standard method to graphically represent the data is Principal Coordinate Analysis (PCoA). It is a technique that helps to extract and visualize informative components of variation from complex, multidimensional data. Basically the goal of PCoA is to map samples from a distance matrix in a set of orthogonal axes. The principal coordinates can be plotted in two or three dimensions. Each of the samples will be visualized explaining the variability between every sample and within a category. Multiple PCoA graphs were generated as output, in order to find possible associations (express by clusters) that can be useful to address biological questions.

5.3.6 Statistical analysis

To assess the composition and diversity within a microbial community, several statistical models have been identified to be suitable in the context of metagenomics data evaluation.

Usually the first thing to be considered in the microbial profile is the abundance count. In metagenomics, it is an integer (0 or positive) that reports the number of taxon identified. Basically it represents the number of times a taxon is identified. For that purpose, the first statistical analysis is computed using the OTU table. Being this type of file a biological matrix reporting samples and associated taxa with every single frequency, it is possible to use various approaches to identify statistical significance within the sample and by different categories. Normalization plays a central role, since it is used as a mean to mitigate the contribution of non-experimental variables in order to minimize their contribution to observed trends. Different variables may affect the differences in the distribution (i.e. healthy versus affected, sample mean, locations). Those variables need to be considered in order to remove variability that is not under experimental control. For that reason, the meta data file is necessary as initial input. In fact, it contains information that link

the variables by different categories. We could be interested in the statistical evaluation of certain identified species, depending on different variables. For example, it is possible to identify taxa that are abundant in a specific subcategory of the samples compared to another (i.e. affected patients versus healthy controls, female versus male or young versus old).

To determine the statistical significance of each taxon assigned, in both studies I used the Analysis of Variance (ANOVA) test [129], together with a Bonferroni correction [131], computed on the OTU tables. For the diversity analysis statistical evaluation I used the ANOSIM [132] and ADONIS [133] tests. Both the tests were used independently on the same data set in order to see if they agree.

CHAPTER 6

RESULTS

6.1 Sequencing results

The next generation sequencing experiment, performed on a 454+ Titanium platform (Roche), generated 730.257 total raw sequences (seq) for the celiac Disease experiment and 40.621 total raw sequences for the Crohn's Disease study.

After the quality filtering step a total of 30.351 sequences, associated with 16S rRNA, were obtained for the Crohn's Disease study, distributed as:

Crohn's affected patient / 27.831 seq, Healthy Control / 22.260 seq, and Patient after nutritional therapy / 8.861 seq. Concerning the OTU picking procedure, 705, 1.328 and 2.171 OTUs were identified, in the patient before and after therapy, and in the control subject, respectively.

The celiac Disease dataset instead, obtained a total of 583.520 post quality filtered sequences associated with 16S rRNA and ITS. Specifically, by 16S bacterial rRNA sequencing I preliminarily obtained 214.999 post quality filtered sequences. After the OTUs picking procedure I identified a total of 2.399 Operational taxonomic units (OTUs). By ITS fungal rRNA sequencing I preliminarily obtained 368.521 post quality filtered sequences. After the OTUs picking procedure a total of 696 Operational taxonomic units were identified. (OTUs).

A sequences summary for both the studies is reported in Table 8.

Table 8. Summary of quality filtering and OTUs picking results for the celiac Disease and Crohn's Disease study.

	Celiac Disease	Crohn's Disease
Number of raw sequences	730.257	40.621
Sequences outside the length of <200 >1000 nucleotides	26.586	1.089
Low quality sequences (Phred < 25)	2.462	551
Mean sequences length (nt)	508	511
Post quality filtered sequences	583.520	30.351
16S rRNA associated sequences	214.999	30.351
ITS associated sequences	368.521	n/a
16S OTUs	2.399	4.204
ITS OTUs	696	n/a

6.2 Taxonomic classification

6.2.1 16S rRNA bacteria profile

Four main phylogenetic levels characterized the ileum microbiome of the Crohn's disease samples: *Bacteriodetes, Firmicutes, Fusobacteria, Proteobacteria.* As shown in Figure 3, *Proteobacteria* were more abundant, and *Bacteroidetes* less abundant in the Crohn's disease patients before therapy (PATIENTS–BT) than in the control. Interestingly, after nutritional therapy (PATIENT–AT) the composition of the ileum microbiome in the patient was virtually the same as in the control. The *Fusobacteria* phylum was present only in the control subject. The Family level classification showed interesting differences between the 3 subjects. Particularly, the *Bacteroidacae* family, belong to the *Bacteriodetes* phylum was dramatically low in patients before therapy, compared to control and patient after therapy. In opposite, the patient before nutritional therapy reported the highest number of all the taxa in the Crohn's study belongs to the *Enterobacteriacae* family (*Proteobacteria* phylum). In fact, in both control and patient after therapy this family was 5 times less represented. Other minor differences have been reported. An overall taxonomic profile at family level is shown in the heatmap table of Figure 4. The heatmap was obtained filtering the OTUs using a threshold of 100 seq/sample. In brownish red are highlighted the most significant alterations in the bacterial composition of gut microbiome detected in the Cronh's disease patient before therapy. Every number in the heatmap indicates the number of sequences associated with a specific bacterial family.

The taxonomic composition for the celiac disease (CD) study was computed from 2.399 Operational taxonomic units (OTUs) identified. From all the OTUs, ~0,5% were unclassified and 95% known bacteria after the taxonomic classification (Figure 5).

Seven main phyla were identified: *Actinobacteria* (9,6%), *Bacteroidetes* (16,9%), *Cyanobacteria* (0,8%), *Firmicutes* (18,6%), *Fusobacteria* (6,8%), *Proteobacteria* (45,6%), *Spirochaetes* (1%). The analysis of the sequences showed also a 0,5% of unclassified bacteria phyla and 217 sequences of which only the phylum could be assigned. These latter sequences were mainly assigned to the *Firmicutes, Proteobacteria* and to *TM7* phyla and interestingly, they were present twice in CD-Patients than in Controls or in gluten diet free patients (GDF-Patients). After filtering the OTUs with a minimum of 10 sequences per OTU, 33 Classes, 56 Orders, 98 Families, and 170 different genera of known bacteria were identified. A significant difference between CD-Patients and Controls subjects was observed at the class level in *Betaproteobacteria* ($p = 0.005$, ANOVA).

The Order level comparison among groups highlighted a statistical significance difference ($p = 0.048$, ANOVA) in the order *Neisseriales, which* abounded in the active CD-Patients (19%) respect to GFD-Patients (2%) and Controls (6%), (Figure 6). I observed that this difference remained significant between CD-Patients and Controls at genus level only for *Neisseria* ($p = 0.008$, after Bonferroni correction). Furthermore, *Neisseria* is the most represented (99,8%) over all the genera associated to the order *Neisseriales* (*Neisseria, Conchiformibius, Elkenella, Kingella*). The majority of known bacteria sequences were classified within five genera: *Acinetobacter* (12,5%), *Sreptococcus* (10,6%), *Haemophilus* (9,2%), *Prevotella* (8,6%) and *Neisseria* (8,5%), (Table 9). After filtering the OTUs with a minimum of 500 sequences per OTU, 12 most represented species were identified, while for other 9 taxa the taxonomic assignment was not able to associate any bacteria species (Table 9). The identified species *Parainfluenzae* belong to the *Haemophilus* family was the most represented in the Controls and in CD-Patients, while it was ten times less present in GFD-Patients, in which less species have been identified compared to the other groups (Table 9).

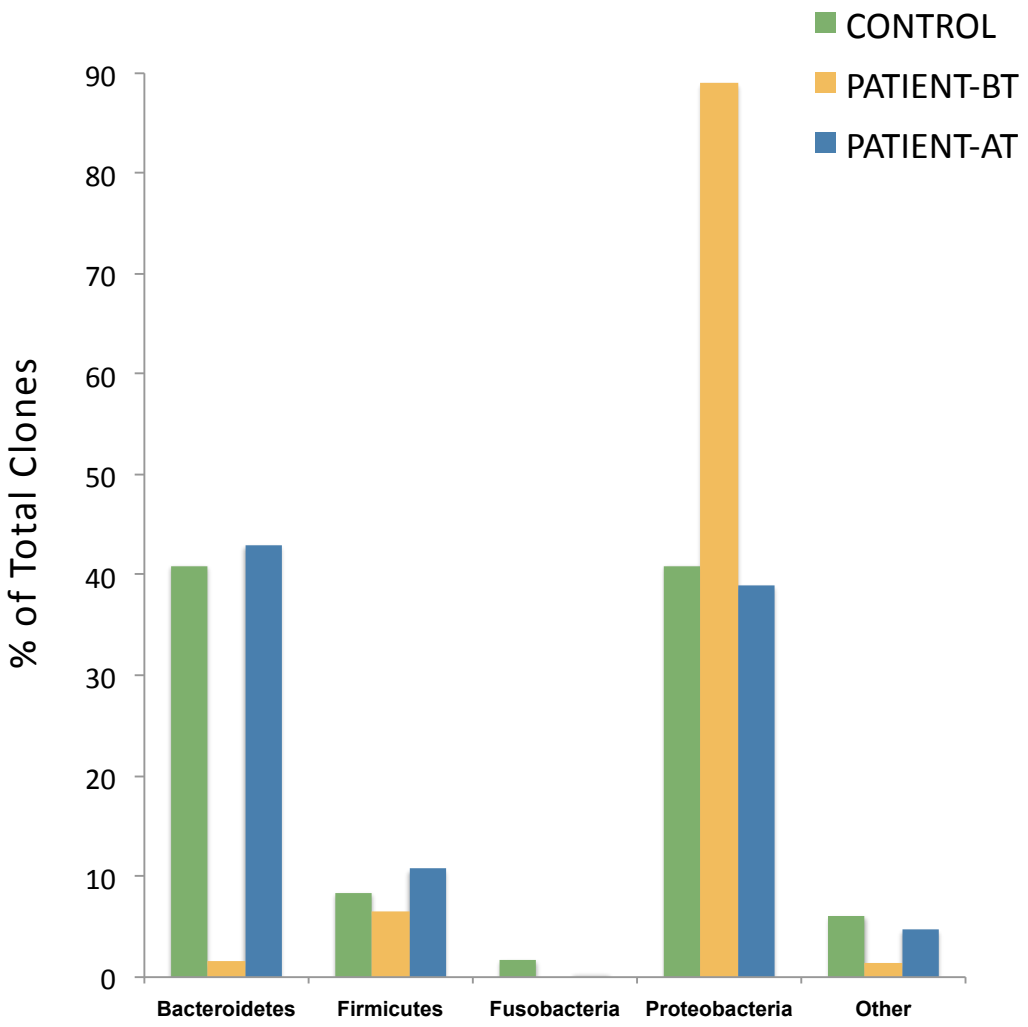

Figure 3. Phylum composition of Crohn's disease study.

Composition of the ileum microbiome characterized in the control subject and in the Crohn's disease patient before therapy (patient-BT) and after therapy (patient-AT) by next-generation sequencing of the 16S rRNAs. *Proteobacteria* were more abundant and *Bacteroidetes* less abundant in the Crohn's disease patient before therapy (patient–BT) than in the control. The *Fusobacteria* phylum is present only in the control subject. After nutritional therapy the composition of the ileum microbiome in the patient after therapy (patient-AT) was virtually the same as in the control.

Phylum	Class	Order	Family	Control	Patient-BT	Patient -AT	#Number of OTU ID
Proteobacteria	Betaproteobacteria	Burkholderiales	Alcaligenaceae	158	0	0	1
Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	2625	138	4908	5
Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	1294	7610	2639	10
Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	5	56	48	1
Firmicutes	Clostridia	n\a	n\a		144	210	1
Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	7	441	6	2
Bacteroidetes	Bacteroidia	Bacteroidales	Porphyromonadaceae	10	0	367	2
Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	106	0	61	1

Figure 4. Heatmap of family level taxonomic classification identified in Crohn's disease study.

The heatmap was obtained filtering the OTUs at a threshold of 100 seq/sample. Every number in the heatmap indicates the number of sequences associated with a specific bacterial family. In brownish red are highlighted the most significant alterations in the bacterial composition of the gut microbiome detected in the Cronh's disease patient before therapy (patient-BT). In green are reported the bacterial families, which are equally distributed between the control and the patient after therapy (patient-AT). In blue are reported all the families equally distributed or with non-significant differences between the three subjects. The *Bacteroidacae* family, belonging to the *Bacteriodetes* phylum, was dramatically low in the patient before therapy, compared to control and patient after therapy. In opposite, the patient before nutritional therapy reported the highest number of all the taxa in the Crohn's study, belonging to the *Enterobacteriacae* family (*Proteobacteria* phylum). The number of OTU ID (last column), refers to the number of genera potentially belonging to a family. *Enterobacteriacae* is the richest family in terms of different bacteria genera than every other identified family.

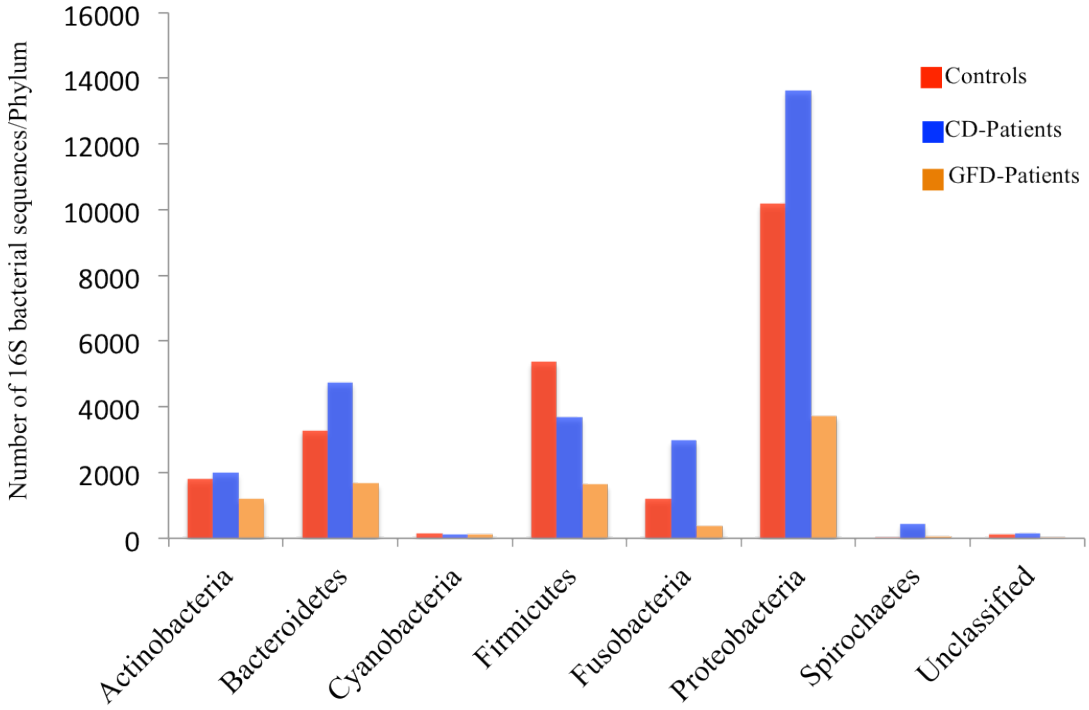

Figure 5. Phylum level classification among the 3 tested groups (Controls, CD-Patients, GFD-Patients) in celiac disease study.

Phylum level classification among the 3 tested groups (Controls, CD-Patients, GFD-Patients) reporting the number of bacterial sequences found. Seven main phyla were identified: *Actinobacteria* (9,6%), *Bacteroidetes* (16,9%), *Cyanobacteria* (0,8%), *Firmicutes* (18,6%), *Fusobacteria* (6,8%), *Proteobacteria* (45,6%), *Spirochaetes* (1%). *Proteobacteria* was the most represented phylum in all the groups while *Cyanobacteria* was less abundant. *Proteobacteria* were highly increased in CD-Patients compared with the other two groups even if no significant differences were found in each of the identified phylum among groups. Unclassified bacteria represented the ~0,5% in the average of every group.

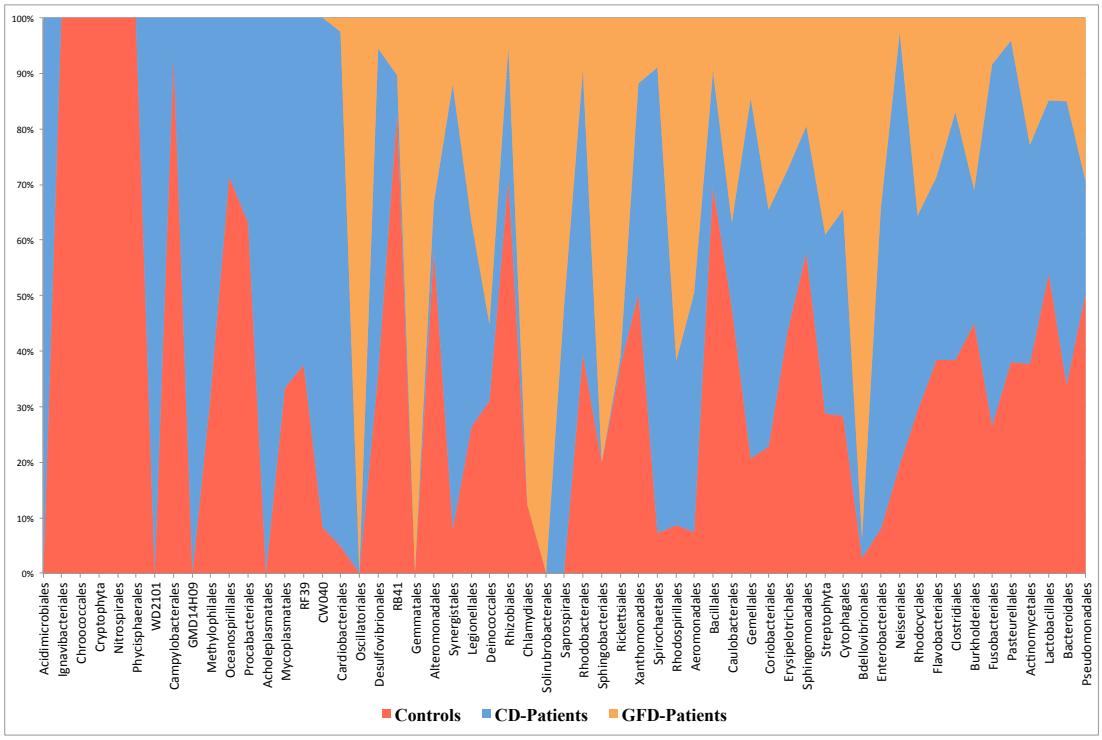

Figure 6. Order level comparison among groups in celiac disease.

The Order level comparison highlighted the presence of 56 different orders identified among the 3 groups. *Pasteurellales* was the most abundant in CD-Patients and second most abundant in controls after *Lactobacillales*. The order *Pseudomonadales* was the most abundant in GFD-Patients and controls. The comparison among groups highlighted a statistical significance difference ($p = 0.048$, ANOVA) in the order *Neisseriales,* which abounded in the active CD-Patients (19%) compared to GFD-Patients (2%) and Controls (6%). Inside the *Proteobacteria* phylum, was observed a trend in reduction for the *Pseudomonadales* order (*Gammaproteobacteria class*) in active-CD Patients respect to the other groups, which was significant at family level for *Pseudomonacae* ($p = 0.002$) between active-CD and GFD Patients.

Table 9. Species level taxa count among groups in celiac disease.

Phylum	Class	Order	Family	Genus	Species	Controls	CD-Patients	GFD-Patients
Actinobacteria	Actinobacteria	Actinomycetales	Micrococcaceae	Rothia	Mucila-ginosa	83	514	21
Actinobacteria	Actinobacteria	Actinomycetales	Propionibacteriaceae	Propionibacterium	acnes	389	647	529
Bacteroidetes	Bacteroidia	Bacteroidales	Paraprevotellaceae	Prevotella	n\a	306	463	149
Bacteroidetes	Bacteroidia	Bacteroidales	Porphyromonadaceae	Porphyromonas	n\a	100	418	50
Bacteroidetes	Bacteroidia	Bacteroidales	Porphyromonadaceae	Porphyromonas	Endo-dontalis	51	223	60
Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella	intermedia	51	301	19
Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella	Melanino-genica	846	698	447
Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella	nanceiensis	59	247	24
Firmicutes	Bacilli	Gemellales	Gemellaceae	n\a	n\a	81	262	80
Firmicutes	Bacilli	Lactobacillales	Carnobacteriaceae	Granulicatella	n\a	215	158	223
Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus	n\a	1368	827	285
Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus	infantis	1098	551	147
Firmicutes	Erysipelotrichi	Erysipelotrichales	Erysipelotrichaceae	Bulleidia	moorei	153	108	121
Fusobacteria	Fusobacteriia	Fusobacteriales	Fusobacteriaceae	Fusobacterium	n\a	939	2268	174
Proteobacteria	Beta-proteobacteria	Neisseriales	Neisseriaceae	Neisseria	all others	793	2613	84
Proteobacteria	Beta-proteobacteria	Neisseriales	Neisseriaceae	Neisseria	subflava	294	1752	49
Proteobacteria	Gamma-proteobacteria	Pasteurellales	Pasteurellaceae	Actinobacillus	Parahaemo-lyticus	288	744	8
Proteobacteria	Gamma-proteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus	n\a	227	535	19
Proteobacteria	Gamma-proteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus	Para-influenzae	2194	2339	270
Proteobacteria	Gamma-proteobacteria	Pseudomonadales	Moraxellaceae	Acinetobacter	n\a	2512	785	1551
Proteobacteria	Gamma-proteobacteria	Pseudomonadales	Moraxellaceae	Acinetobacter	johnsonii	206	181	50

6.2.2 ITS fungal profile in celiac disease study

By ITS fungal sequences, only two main phyla were identified in the celiac disease study: *Ascomycota* (50%) and *Basidomycota* (44%). There were no statistical differences in the phylum level abundance in the 3 tested groups.
A 6% of sequences belonged to unidentified or uncultured fungi (Figure 7). Even if the ANOVA test reported a significant difference ($p < 0.05$) for the order *Malasseziales*, belonging to the *Basidomycota* phylum, after Bonferroni correction the value was not statistically significant.
At the family level, a total of 36 families were identified, of which 20 belong to the *Ascomycota* phylum and 16 families to the *Basidomycota* (Figure 8). The family *Mycosphaerellaceae* was the most represented in the three groups, even though no statistical difference was found.
Using a filter of 200 seq/sample a total of 46 genera were identified in the two phyla *Ascomycota* and *Basidiomycota* (Table 10). The most represented genera were *Cladosporiumin* and *Candida* in the *Ascomycota* and *Cryptococcus* in the *Basidiomycota*. In particular, the *Candida* genus was more abundant and *Cryptococcus* less abundant in CD-Patients than in the other two groups, even if at not statistically significant level.

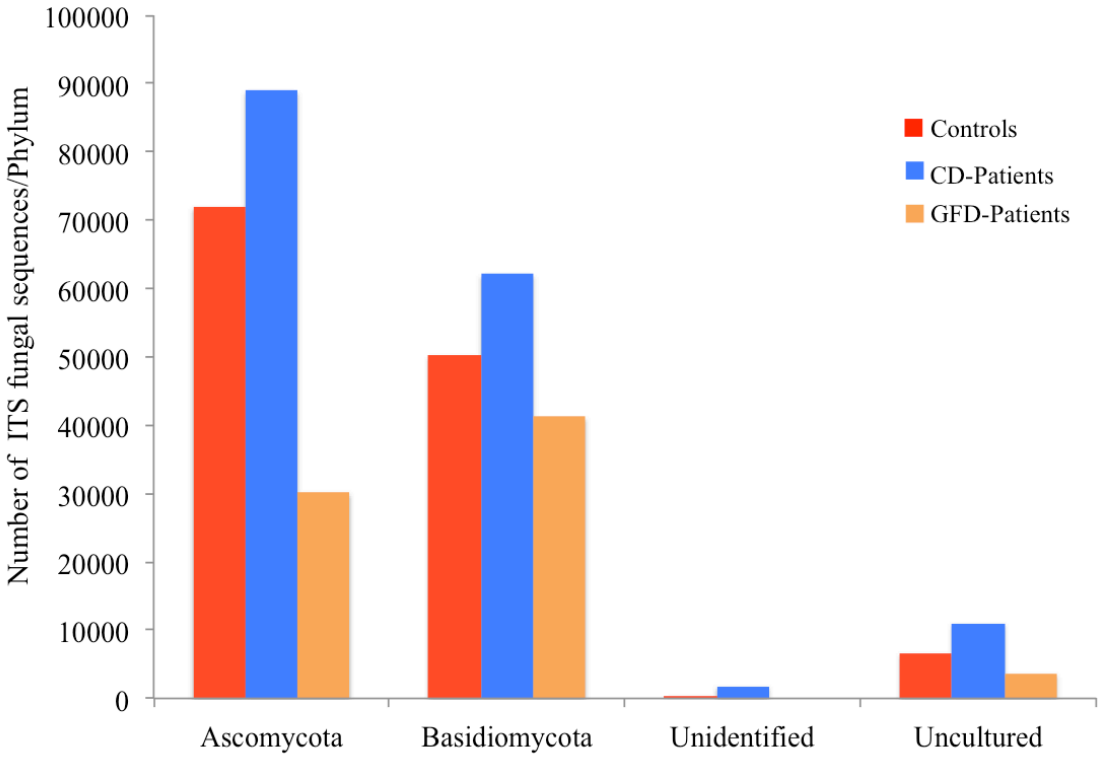

Figure 7. ITS phylum level classification in celiac disease study.

Two main fungal phyla were identified in the 3 groups of studied subjects. *Ascomycota* (50%), *Basidiomycota* (44%), *Unidentified* (0,5%), *Uncultured* (5,5%). Uncultured are those ITS that have been identified but not yet annotated. Unidentified are those ITS that have not been yet identified and not yet annotated. The latter ITS represents the 1% of the total ITS number in CD-Patients, 0,2% in the controls group and 0% in GFD-Patients. *Ascomycota* phylum was less abundant in GFD-Patients compared to the other groups, but there were no statistical differences in the phylum level abundance in the 3 tested groups.

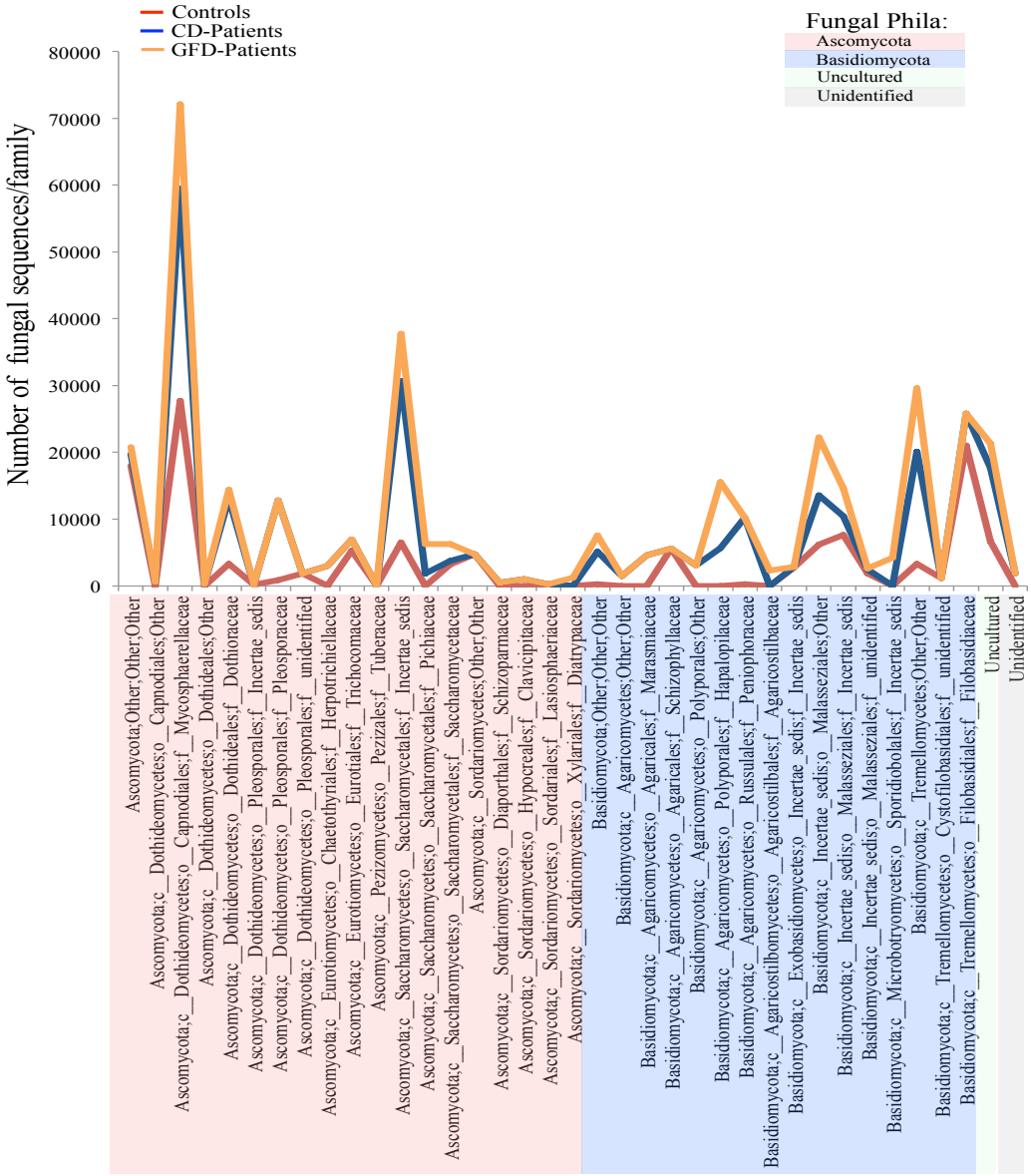

Figure 8. ITS family level classification in celiac disease study.

A total of 35 families were identified within the *Ascomycota* and *Basidiomycota* phyla. The most represented were *Mycosphaerellaceae* and *Incertae sedis* family belonging to *Capnodiales* and *Saccharomycetales* order respectively (both within the *Ascomycota* phylum), which were both more abundant in active-CD and GFD-Patients than in the controls.

Table 10. ITS genus level classification in celiac disease.

Phylum	Class	Order	Family	Genus	Controls	CD - Patients	GFD-Patients
Ascomycota	Dothideomycetes	Capnodiales	Other	Other	0	665	0
Ascomycota	Dothideomycetes	Capnodiales	Mycosphaerellaceae	Other	1338	4995	3
Ascomycota	Dothideomycetes	Capnodiales	Mycosphaerellaceae	Cladosporium	26396	24763	12614
Ascomycota	Dothideomycetes	Capnodiales	Mycosphaerellaceae	Ramularia	0	1981	0
Ascomycota	Dothideomycetes	Dothideales	Other	Other	0	308	0
Ascomycota	Dothideomycetes	Dothideales	Dothioraceae	Aureobasidium	3287	9831	1265
Ascomycota	Dothideomycetes	Pleosporales	Pleosporaceae	Alternaria	895	7446	0
Ascomycota	Dothideomycetes	Pleosporales	Pleosporaceae	Epicoccum	0	4461	0
Ascomycota	Dothideomycetes	Pleosporales	unidentified	unidentified	1874	0	0
Ascomycota	Eurotiomycetes	Chaetothyriales	Herpotrichiellaceae	unidentified	0	2967	3
Ascomycota	Eurotiomycetes	Eurotiales	Trichocomaceae	Penicillium	5282	1652	0
Ascomycota	Pezizomycetes	Pezizales	Tuberaceae	Choiromyces	59	0	0
Ascomycota	Saccharomycetes	Saccharo-mycetales	Incertae_sedis	Candida	6471	24209	7045
Ascomycota	Saccharomycetes	Saccharo-mycetales	Pichiaceae	Pichia	2	1854	4400
Ascomycota	Saccharomycetes	Saccharo-mycetales	Saccharomycetaceae	Debaryomyces	3262	539	2447
Ascomycota	Sordariomycetes	Other	Other	Other	4726	0	0
Ascomycota	Sordariomycetes	Diaporthales	Schizoparmaceae	Coniella	0	473	0
Ascomycota	Sordariomycetes	Hypocreales	Clavicipitaceae	Other	0	1045	0
Ascomycota	Sordariomycetes	Sordariales	Lasiosphaeriaceae	Other	196	0	0
Ascomycota	Sordariomycetes	Xylariales	Diatrypaceae	Other	4	0	1252
Ascomycota	Sordariomycetes	Xylariales	Diatrypaceae	Eutypa	0	0	34
Basidiomycota	Agaricomycetes	Other	Other	Other	0	1484	0
Basidiomycota	Agaricomycetes	Agaricales	Marasmiaceae	Hemimycena	0	4617	0
Basidiomycota	Agaricomycetes	Agaricales	Schizophyllaceae	Schizophyllum	5598	0	0
Basidiomycota	Agaricomycetes	Polyporales	Other	Other	0	3052	0
Basidiomycota	Agaricomycetes	Polyporales	Hapalopilaceae	Bjerkandera	1	0	9849
Basidiomycota	Agaricomycetes	Polyporales	Hapalopilaceae	Ceriporiopsis	1	5691	0
Basidiomycota	Agaricomycetes	Russulales	Peniophoraceae	Peniophora	265	9979	0
Basidiomycota	Agaricostilbomycetes	Agaricostilbales	Agaricostilbaceae	Bensingtonia	0	0	2327
Basidiomycota	Exobasidiomycetes	Incertae_sedis	Incertae_sedis	Tilletiopsis	2814	29	0
Basidiomycota	Incertae_sedis	Malasseziales	Other	Other	6232	7353	8637
Basidiomycota	Incertae_sedis	Malasseziales	Incertae_sedis	Malassezia	7636	2850	4115
Basidiomycota	Incertae_sedis	Malasseziales	unidentified	unidentified	1864	582	170
Basidiomycota	Microbotryomycetes	Sporidiobolales	Incertae_sedis	Sporobolomyces	5	0	4199
Basidiomycota	Tremellomycetes	Other	Other	Other	3286	16808	9513
Basidiomycota	Tremellomycetes	Cysto-filobasidiales	unidentified	unidentified	1199	0	2
Basidiomycota	Tremellomycetes	Filobasidiales	Filobasidiaceae	Cryptococcus	21013	4809	14

6.3 Diversity Analysis

6.3.1 Alpha diversity analysis in Crohn's disease study

The alpha diversity was computed using Observed Species method [121], Chao1 estimator [122], the Shannon Diversity Index [123] and the Faith's Phylogenetic Diversity (PD), [124]. The data support the reduced bacterial diversity as computed with all the four different methods for the Crohn's disease patient before therapy than after therapy and even when compared with the control subject. In fact, every method reported a significant bacterial diversity reduction in the Crohn's disease patient before therapy ($p < 0.005$). The observed species metric (Figure 9), at a sequence depth of 10.000 sequences, reported a higher number of different 16S rRNA bacteria species in the control subject, a medium species diversity in the Crohn's disease patient after nutritional therapy (Patient-AT) and the lowest bacteria diversity in the patient before nutritional therapy (Patient-BT).

The alpha diversity computed with Chao1 estimator reported similar results with a slight protrusion in the control subject curve at a rarefaction depth of 4.000 sequences per sample (Figure 10).

A phylogenetic metric, Faith's Phylogenetic Diversity (PD), was also used to compare the alpha diversity of the three samples, considering the phylogenetic distance found in the 16S rRNA sequences. The trend of the curves are very similar to those generated with observed species metric (Figure 11).

Also, a Shannon Diversity Index was computed and the mean scores confirm the community richness within samples to be the highest for the control subject and the lowest for the patient before nutritional therapy. Asterisks refer to statistical significant differences between samples ($p < 0.005$), error bars represent the standard error (Figure 12).

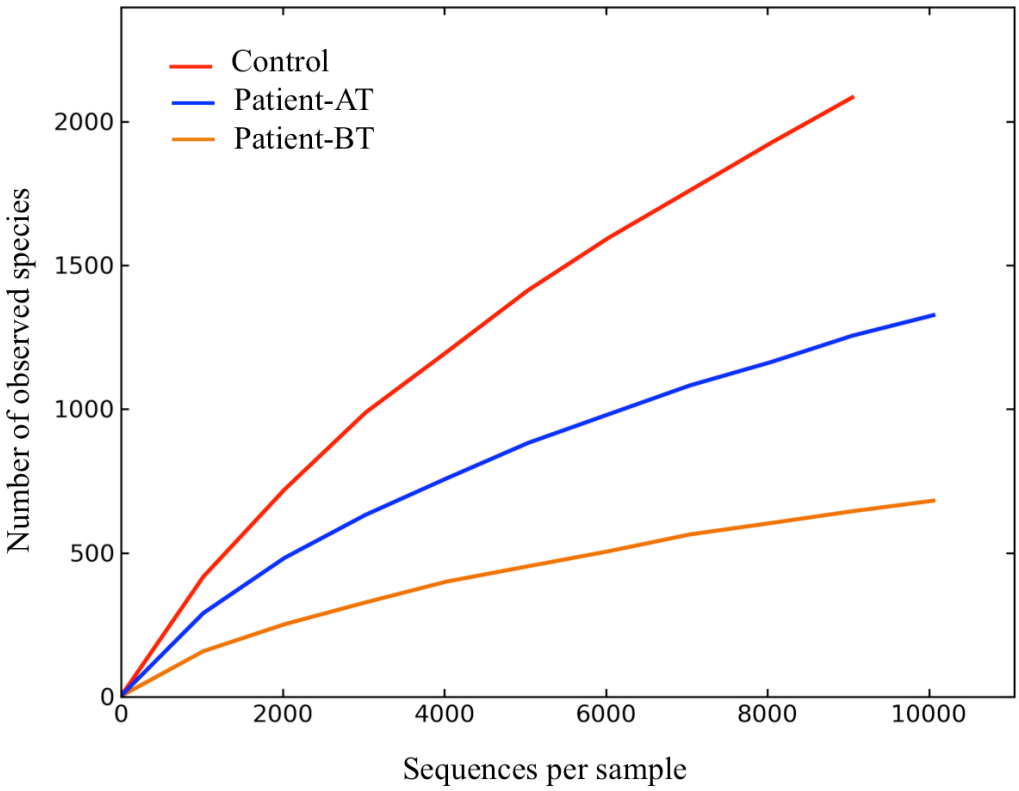

Figure 9. Alpha diversity as observed species in Crohn's disease.

Alpha diversity of 16S rRNAs OTUs as measured using observed species method shows at a sequence depth of 10.000 sequences per sample, a higher number of different 16S rRNA bacteria species in the control subject, a medium species diversity in the Crohn's disease patient after nutritional therapy (Patient–AT) and the lowest bacteria diversity in the patient before nutritional therapy (Patient-BT). The number of different bacteria species grows proportionally with the number of sequences, especially for the control subject. In opposite, the Patient-BT as deductible from the trend of his curve, would reach a limit in bacteria diversity even increasing the number of sequences exponentially.

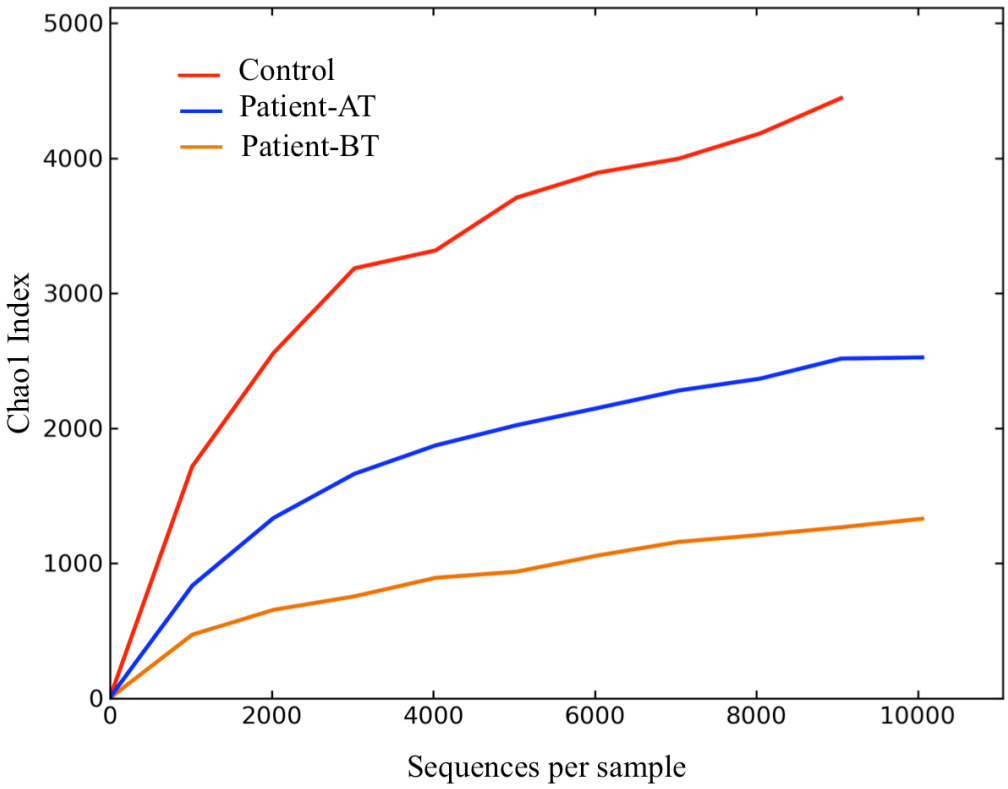

Figure 10. Alpha diversity as Chao1 estimator in Crohn's disease.

The alpha diversity measured using Chao1 estimator shows a slight protrusion in the Control subject curve at a rarefaction depth of 4.000 sequences per sample. This may suggest that over the sampling depth of 4.000 sequences, more rare species are present in the control subject. The other two subjects present a proportional trend in the curves as the number of sequences increases, without showing differences in terms of rare species such as singleton (OTUs with only one sequence) or doubletons (OTUs with only two sequences) as computed in accordance with the Chao1 richness estimator.

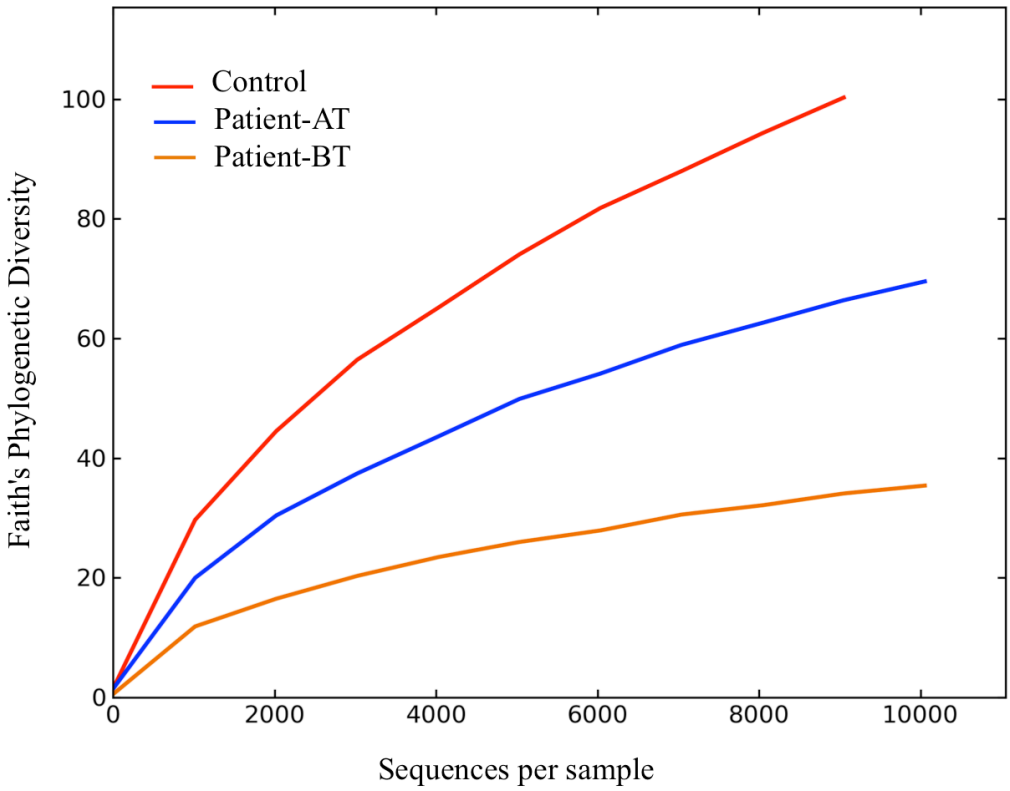

Figure 11. Alpha diversity as Faith's Phylogenetic Diversity (PD) in Crohn's disease.

The alpha diversity was computed using the phylogenetic metric Faith's Phylogenetic diversity. The phylogenetic distances found within the samples confirmed the control subject to have the highest species richness. The Patient-BT presents the lowest bacteria phylogenetic diversity and the Patient-AT is in the middle between the other subjects. The curve belonging to the control seems to present a potentially growth in the phylogenetic bacteria diversity as the number of sequences increases. Same scenario, even with a slower potential growth, occurs for the Patient-AT. The Patient-BT presents the lowest growth of the curve, meaning that species richness computed as phylogenetic distances, would not increase even with a higher number of sequences.

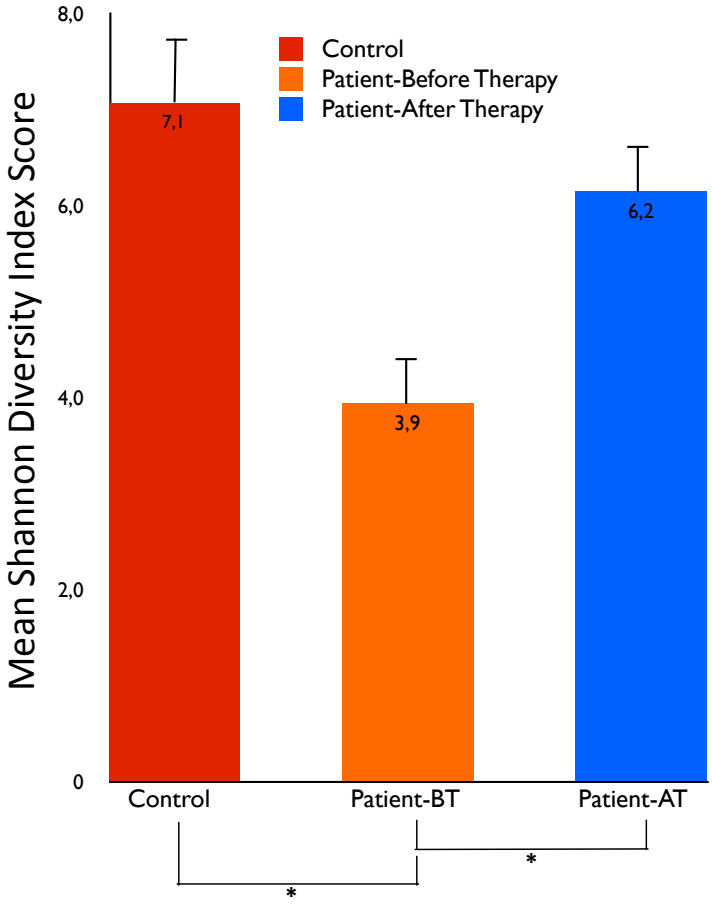

Figure 12. Alpha diversity as mean Shannon Diversity Index Score in Crohn's disease.

The mean Shannon Diversity Index scores confirm the community richness within samples to be the highest for the control subject and the lowest for the patients before nutritional therapy. Asterisks refer to statistical significant differences between samples ($p < 0.005$), error bars represent the standard error.

6.3.2 Alpha diversity analysis in celiac disease study

In the celiac disease study, the alpha diversity analysis was computed for both bacterial and fungal communities. The same metrics used for computing the alpha diversity in Crohn's disease (see 5.3.1), were used for the bacteria and fungal dataset in celiac disease. The only exception concerns the PD metrics, computed only on the bacterial data set (since a template alignment is not yet available for ITS sequences, see 3.3.2).

The 16S rRNA rarefaction curves, at a depth of 420 sequences/sample, showed an equal trend in the number of observed species in all the 3 groups (Figure 13). The Chao1 estimator showed no statistically significant differences in the bacterial community richness, a part a slight lower divergence in the GFD-Patients (Figure 14). The alpha diversity computed with the phylogenetic metric PD, showed an opposite trend for the GFD-Patients, which has the highest curve, while the other two groups showed a similar trend. Also in this case, no statistical significance differences were found. The Shannon Diversity Index Score confirmed that no difference is present in the alpha diversity between the three groups (Figure 16).

The fungal community richness was lower than the one found in the bacterial dataset, at a depth of 2.178 sequences/sample. For both the Chao1 (Figure 17) and observed species (Figure 18) estimators, the Controls and GFD-Patients rarefaction curves showed a very close trend, while the curves obtained for the CD-Patients presented a lower trend. The differences found were statistically significant for both the estimators comparing the alpha diversity of CD-Patients versus the other two groups ($p < 0.05$). No differences were found in the alpha diversity comparing Controls and GFD-Patients ($p = 1$), suggesting that there is lower fungal community richness in the CD-Patients group than Controls and GFD-Patients who reported quite close community richness.

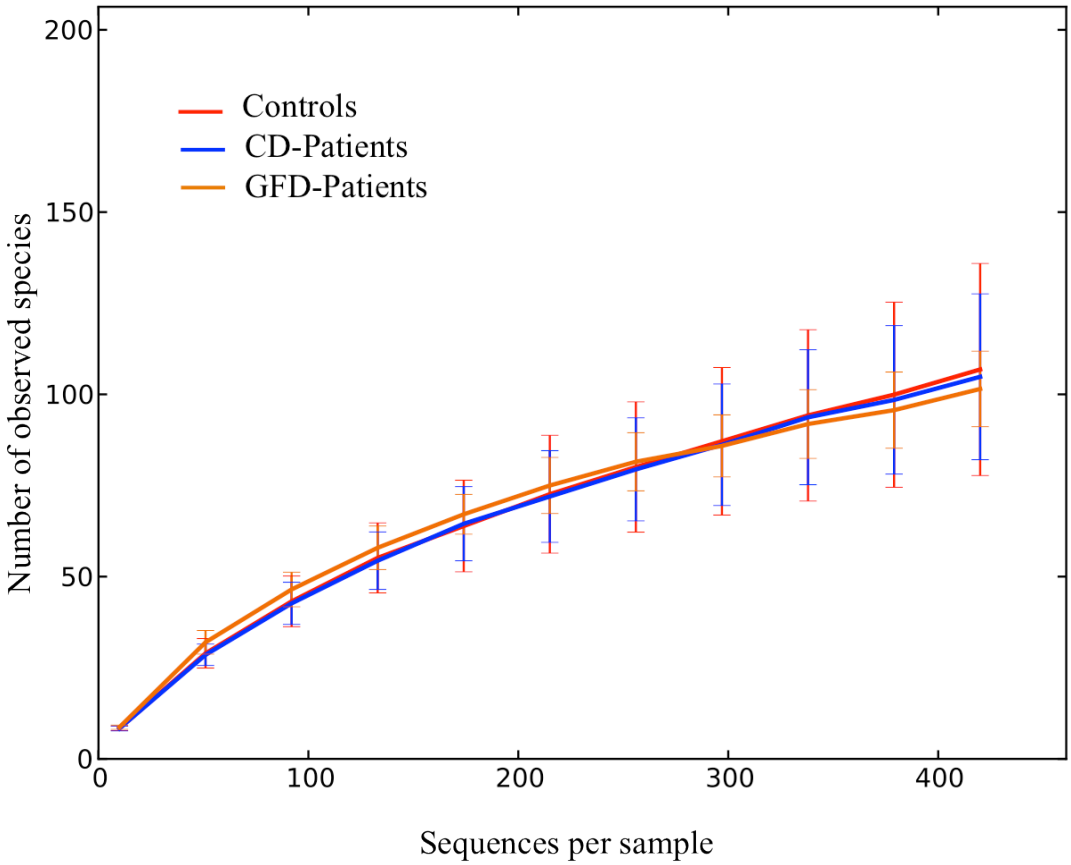

Figure 13. Bacterial alpha diversity as observed species in celiac disease.

The test was computed using a depth of 400 sequences per sample in order to normalize the number of sequences associated to each sample in the study. The rarefaction curves were obtained using observed species method. In this case, the alpha diversity shows no difference among the groups. The trend of each curve is similar, suggesting that the number of different bacteria species increases proportionally with the number of sequences for each sample in the groups of study. Bars represent the standard deviation of the mean at every sequence depth as computed in the rarefaction procedure.

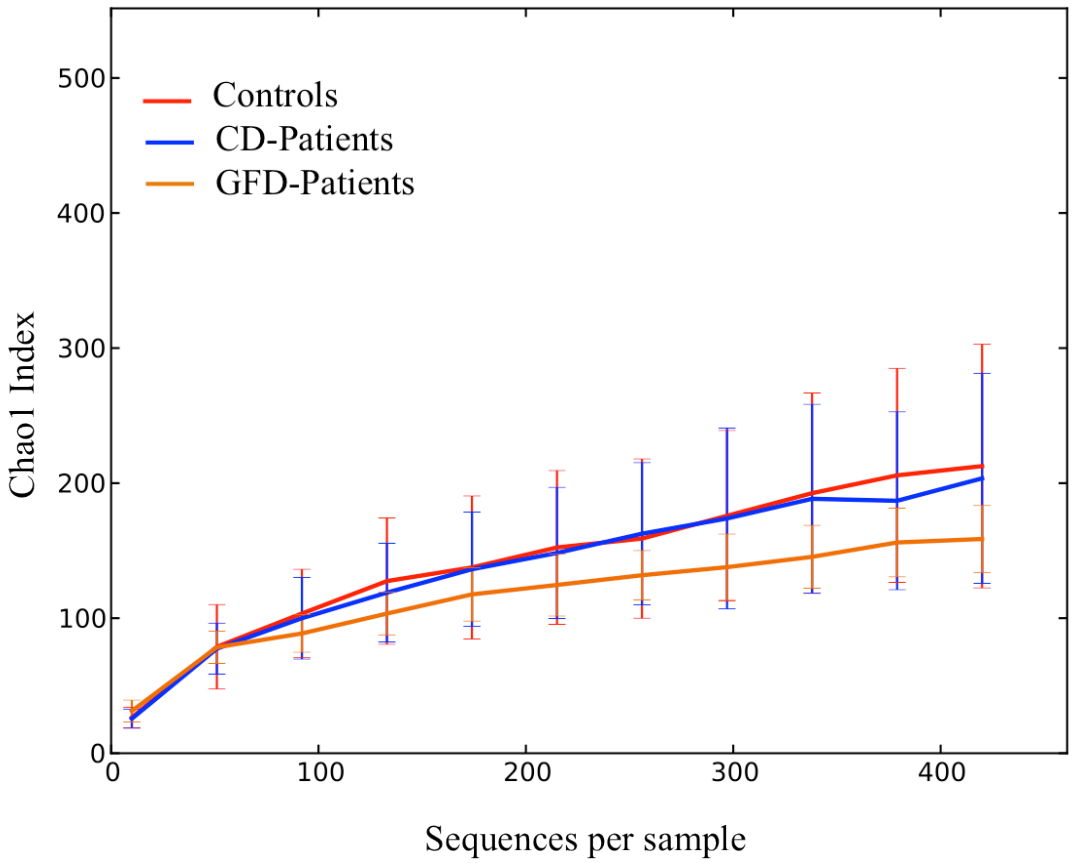

Figure 14. Bacterial alpha diversity as Chao1 estimator in celiac disease.

Alpha diversity of 16S rRNAs OTUs as measured using Chao1 estimator shows a slight even if not statistically significant difference between GFD-Patients and both active-CD patients and controls at a depth of 400 sequences per sample. The result suggests that controls and CD-Patients groups present a proportional trend in the curves as the number of sequences increases, without showing differences in terms of rare species. The curve computed for the GFD-Patients group instead presents a lower number of bacteria species diversity after a depth of 50 sequences/sample. The trend of the curve does not change as the number of sequences increases, meaning that the GFD-Patients group presents a less amount of rare species (singleton and doubletons) compared to the other groups. Bars represent the standard deviation of the mean at every sequence depth as computed in the rarefaction procedure.

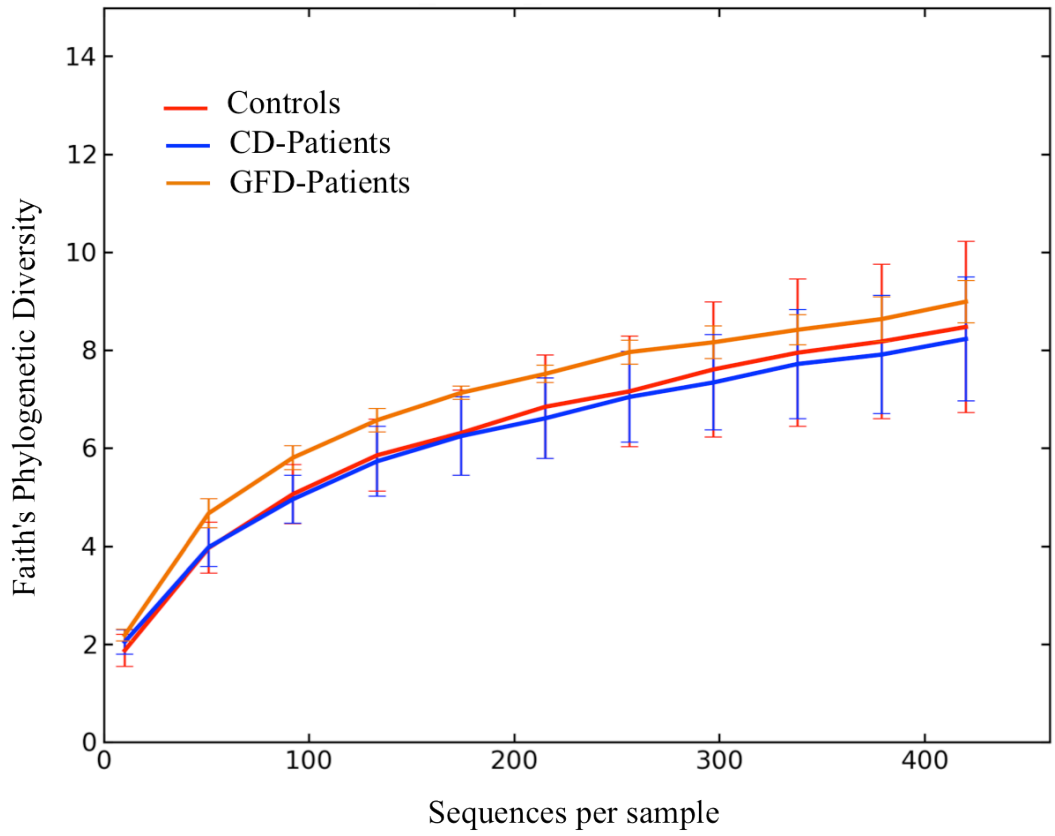

Figure 15. Bacterial alpha diversity as Faith's Phylogenetic Diversity (PD) in celiac disease.

The alpha diversity computed with the Faith's Phylogenetic method reports an opposite result compared to the other methods. In this case, the GFD-Patients group shows the highest curve compared to the other two groups (controls and CD-Patients), which shows no differences. The higher Faith's PD index in GFD-patients group suggests that there is a more, even if not statistically significant, phylogenetic distance in the bacterial species within the GFD-patients group compared to the other two groups. Every curve grows proportionally with the number of sequences, meaning that the difference in the phylogenetic distance related to the identified bacterial species does not depend on the sequences depth. Bars represent the standard deviation of the mean at every sequence depth as computed in the rarefaction procedure.

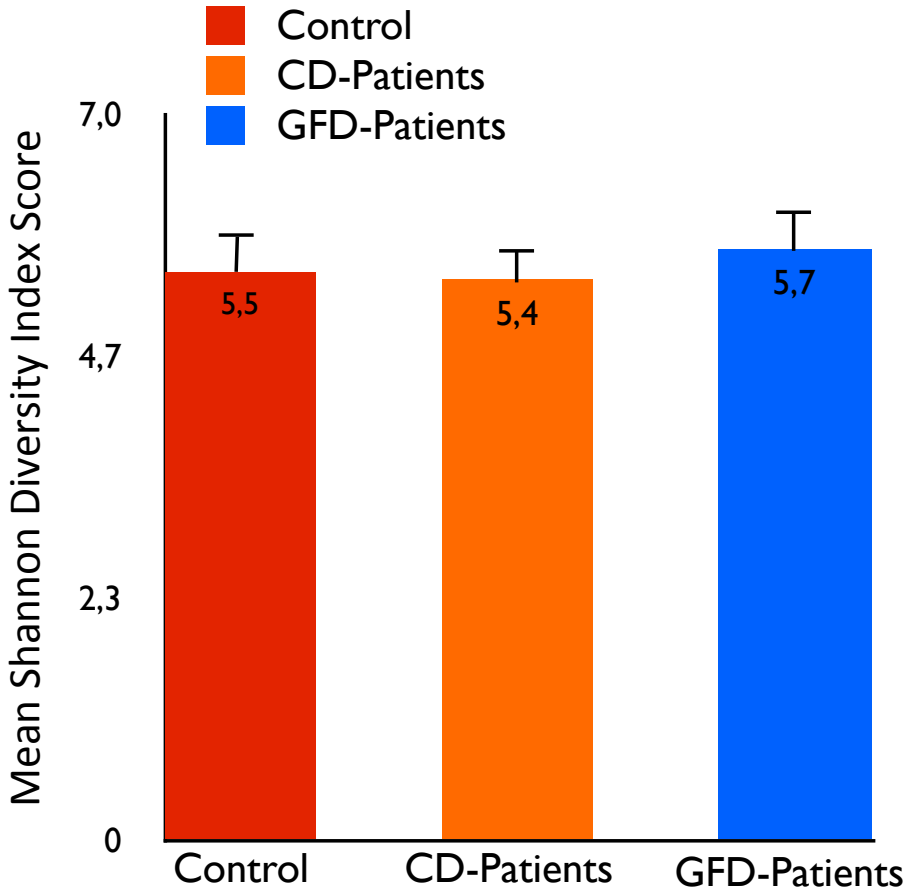

Figure 16. Bacterial alpha diversity as mean Shannon Diversity Index Score in celiac disease.

The mean Shannon Diversity Index scores confirm the community richness within samples to be similar for the three groups of study. No statistical differences were found. Error bars represent the standard error.

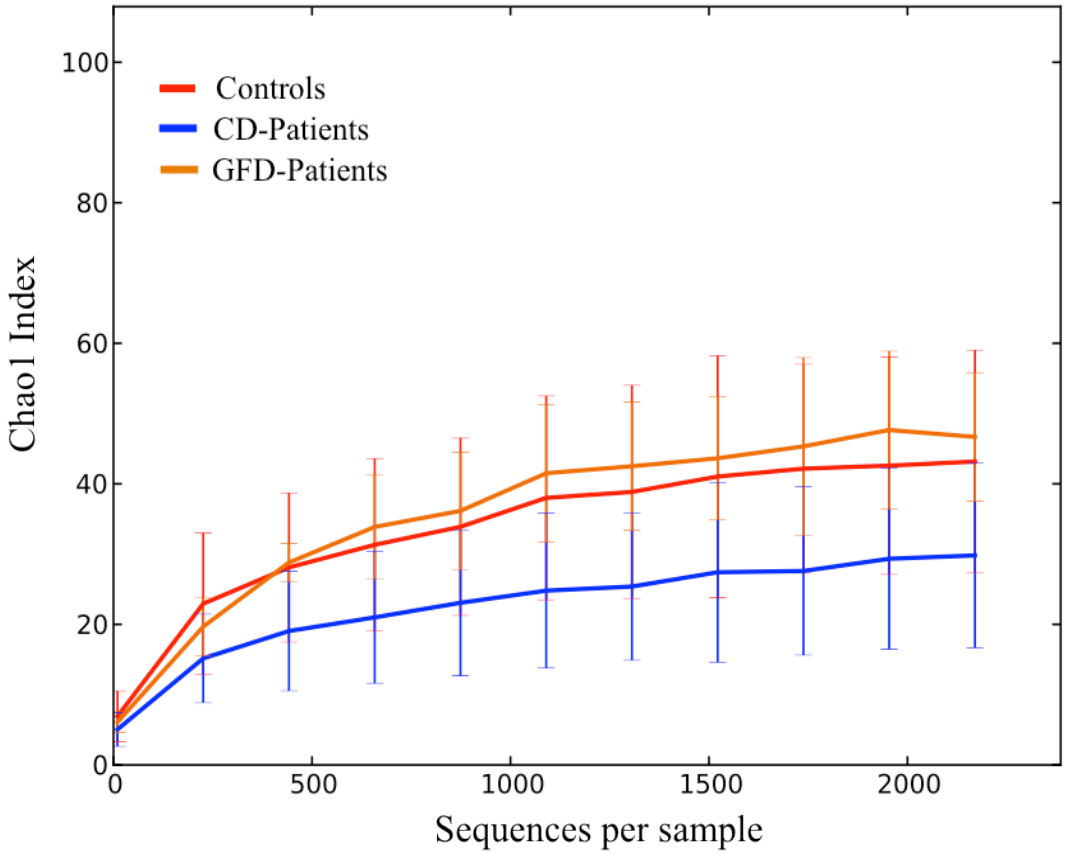

Figure 17. Fungal alpha diversity as Chao1 estimator in celiac disease.

Alpha diversity of fungal OTUs, as measured using Chao1 estimator at a depth of 2.178 sequences per sample, showed a statistically significant difference ($p < 0.05$) between GFD-Patients and both CD-Patients and controls. The CD-Patients group reported a lower fungal diversity compared to controls and GFD-Patients, which curves, have similar trends according to the number of sequences in each sample. Bars represent the standard deviation of the mean at every sequence depth computed in the rarefaction procedure.

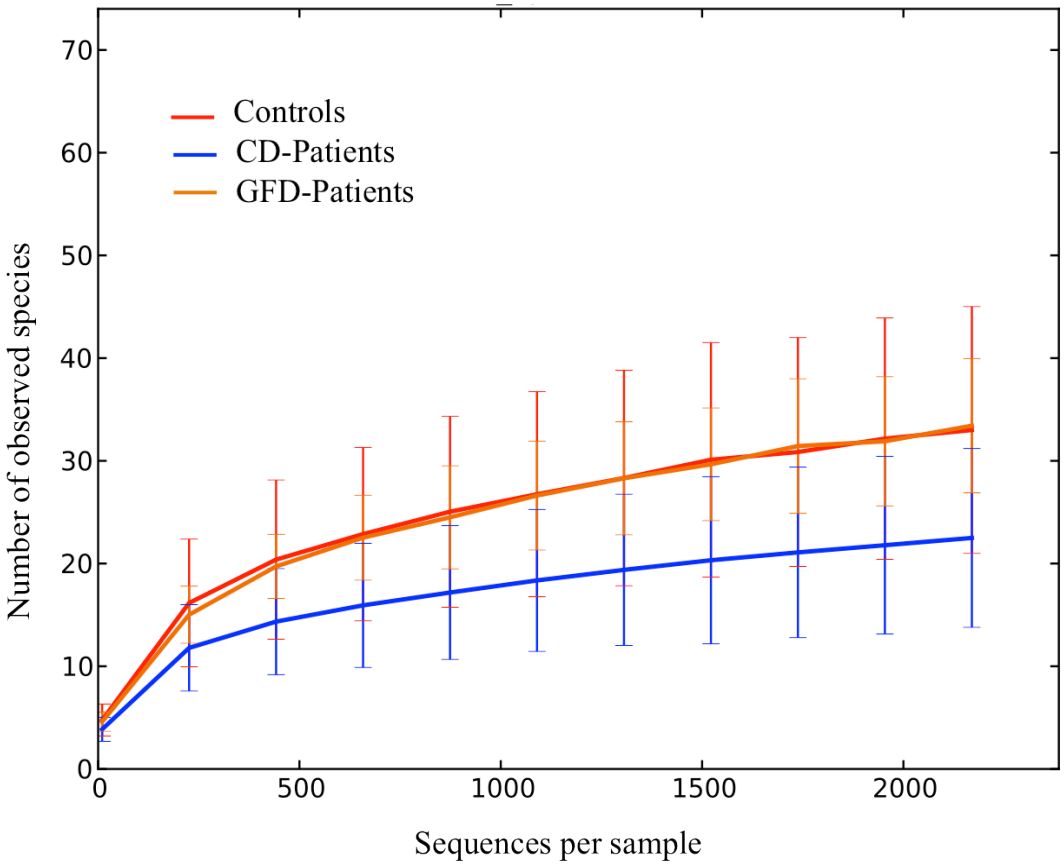

Figure 18. Fungal alpha diversity as observed species in celiac disease.

The alpha diversity with observed species method was performed at a sequence depth of 2.178 sequences per sample. The graph shows an overlapping trend for the curves belonging to the controls and GFD-Patients group. The CD-Patients reported the lowest alpha diversity in terms of different fungal species identified. The curves grow in a similar way as the number of sequences increase, and the differences between GFD-Patients and both CD-Patients and Controls are statistically significant ($p < 0.05$). Error bars represent the standard deviation of the mean at each sequence depth computed during the rarefaction step.

6.3.3 Beta diversity analysis in celiac disease study

The beta diversity, as expression of bacterial community similarity between the samples in the studied groups, was computed only for the celiac disease dataset. In fact, being Crohn's disease study only formed by three samples, the beta diversity would not have been sufficient to highlight any real inter-community similarity.

The weighted and unweighted Unifrac tests were computed at a depth of 420 sequences/sample, in order to explain the beta diversity using a phylogenetic distance matrix. The results are reported in form of PCoA for both the methods. The weighted Unifrac is reported in the PCoA plot (Figure 19) and highlights how the CD-Patients and GFD-Patients subjects cluster in two distinct groups, whereas control subjects showed a random distribution. The only exception in the CD-Patients group cluster, was represented by a 14-year-old CD-Patient that clustered separately (Figure 19, black arrow).

The unweighted Unifrac instead showed no significant clusters between the samples, suggesting that the presence/absence of taxa in the community of each sample is highly influenced by the number of sequences/sample collected.

The fungal beta diversity was computed using Bray-Curtis matrix. The PCoA plot showed no significant clusters within the group of samples, even if a slight cluster can be observed in the center for the CD-Patients group, although with no statistical significance (Figure 21).

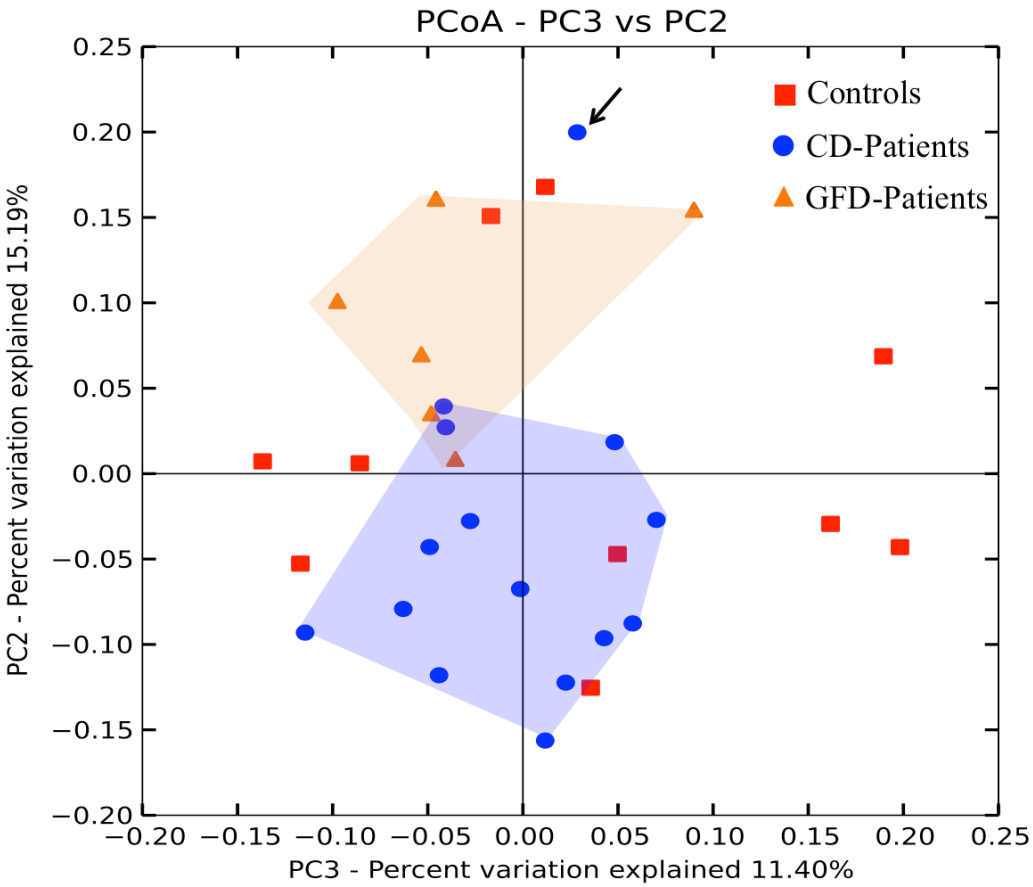

Figure 19. Bacterial beta diversity analysis among samples computed by weighted Unifrac in celiac disease.

Beta diversity analysis among samples computed by weighted Unifrac at a 420 sequences depth. Different clouds are associated with the subgroup of active-CD patients and GFD-patients. A random distribution is observed instead for the control subjects, which did not form any specific cluster. Interestingly, the only active CD-Patients sample that did not cluster within the CD-Patients group is the only 14-year-old female in the CD-patients enrolled in the study (black arrow). Although the weighted Unifrac result shows the presence of clusters, the statistical tests showed a non-significant variation to explain the observed variability, even if the value was close to the significance ($p = 0.09$; ADONIS, ANOSIM). In this case, only few samples highly influence the result, suggesting that a more significant variability could be observed increasing the number of samples in the study.

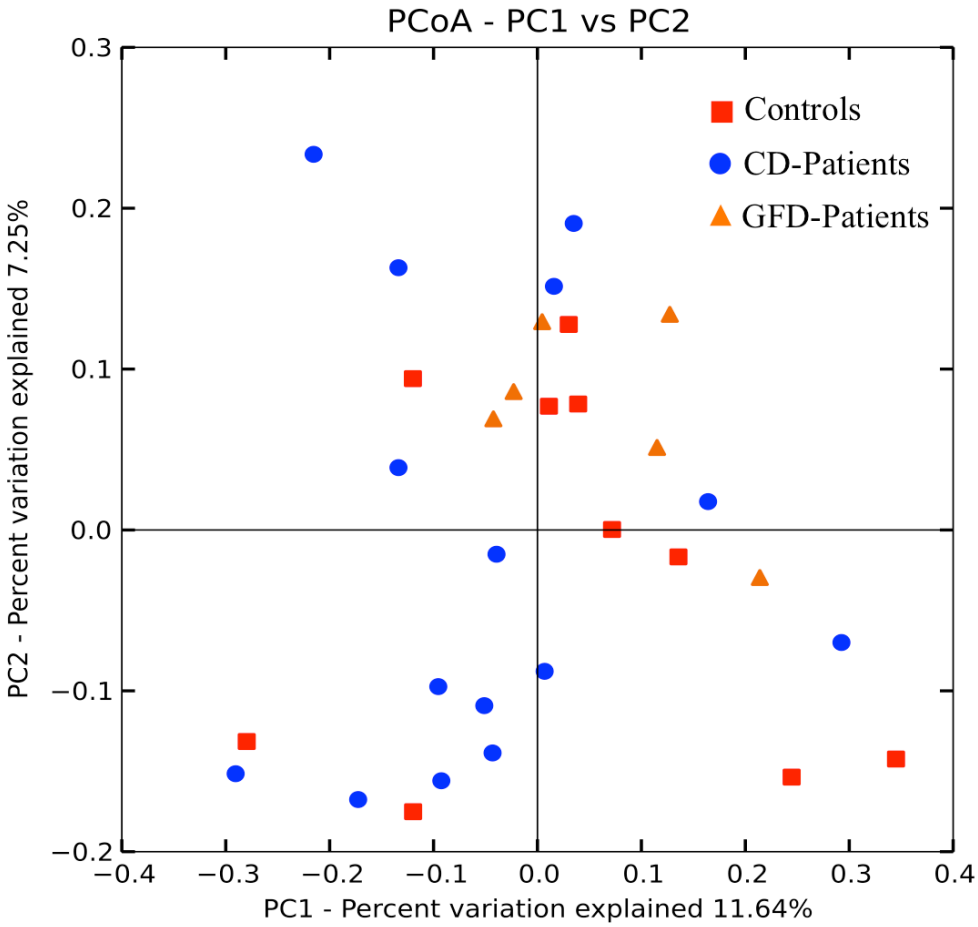

Figure 20. Bacterial beta diversity analysis among samples computed by unweighted Unifrac in celiac disease

The beta diversity computed by non-weighted Unifrac method at a 420 sequences depth, showed a completely random distribution for every sample. In opposite to the weighted Unifrac method, which takes into account the number of sequences, the non-weighted Unifrac generates a qualitative result, only considering the presence/absence of a taxon among samples. Since the observed distribution is random, it might suggest that the number of sequences that belongs to each sample influences the observed microbiome profile. Samples with less sequences do not show the same microbiome profile as those with a higher number of sequences, meaning that the presence of more rare bacteria species is driven by the number of 16S rRNAs amplicon sequences obtained after the next generation sequencing run.

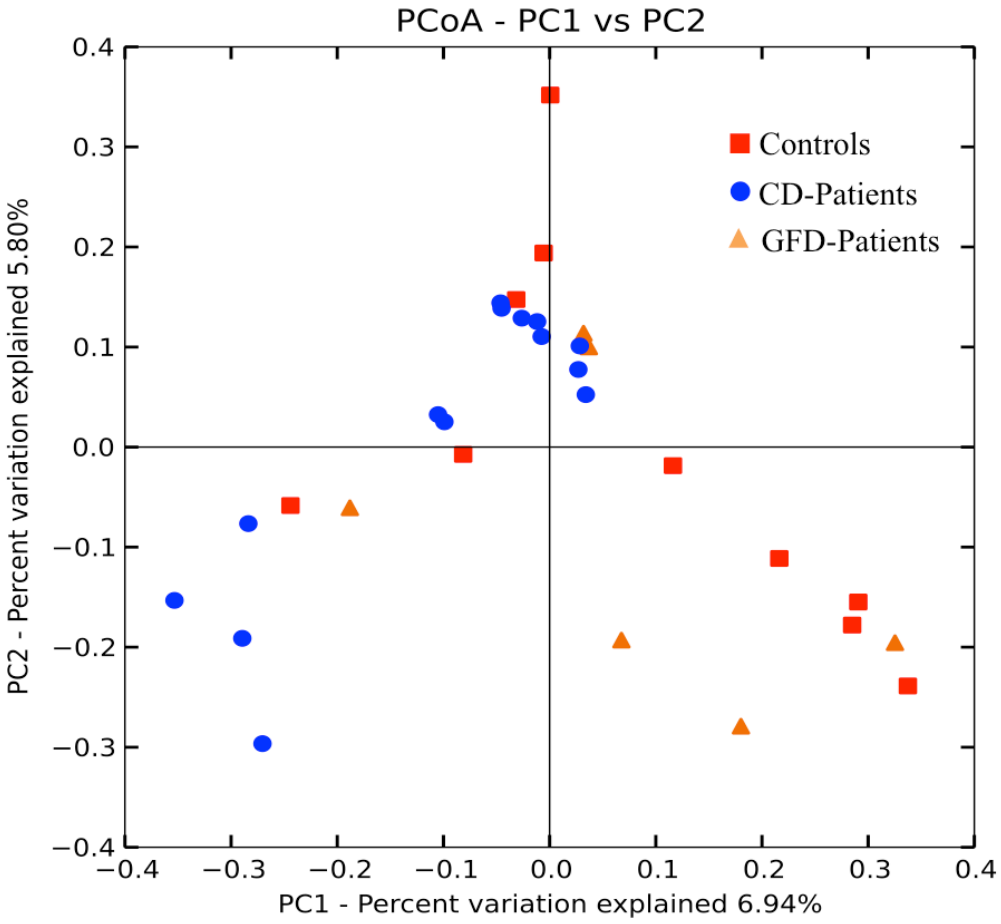

Figure 21. Fungal beta diversity analysis among samples computed by Bray-Curtis in celiac disease.

The fungal beta diversity was computed with Bray-Curtis method at a depth of 2.178 sequences per sample. A slight divergence is shown in the PC1 vs. PC2 in the middle of the plot for the Controls and the CD-Patient groups, even if not significant clusters were reported. A random distribution is instead observed for the GFD-Patients group. The beta diversity between the samples for the fungal community of the gut microbiome does not seem to be related to the type of condition related to the enrolled subjects (healthy controls, active-CD Patients, GFD-Patients).

CHAPTER 7

DISCUSSION

7.1 The altered gut microbiome in a Crohn's disease patient is normalized after nutritional therapy

The human gut hosts one of the most densely populated microbial community on earth, compared to several environments. The number of bacteria exceeds human cells by more than ten-fold and the number of the total bacteria genes holds 100 fold more genes than those of its host [90].

Inflammatory bowel disease (IBD) represents a chronic inflammatory condition of the gastrointestinal tract and is widely associated with the microbial communities of the human gut. In the past few years, different studies have linked IBD with altered interactions between gut microbes and the intestinal immune system. Though, the precise nature of the intestinal microbiota dysfunction in IBD is unclear [145].

Between the IBD a main subtype is of relevant interest due to its high incidence, and is known as Crohn's disease, which includes defined microbial perturbations and tissue localizations. Cronh's disease may affect any part of the digestive tract, and the implications related to the microbial involvement are unclear [146].

In fact, the role of the gut microbiome in Cronh's disease onset and its alteration in the course of active treatment and recovery are still unknown. One of the causes may be correlated to the inability to control bacterial proliferation in the intestinal walls, which may drive microorganisms to take advantage of the host mucosal layer.

The use of different antibiotics targeting different bacterial strains, have highlighted particular taxa in Crohn's disease, suggesting that different pathogens are involved [68]. Current medications designed to treat Crohn's disease are focused in suppressing the abnormal inflammatory response that is the primary cause of the symptoms. The inflammatory suppression offers an initial relief for the Crohn's disease patients, decreasing the symptoms and allowing the intestinal tissues to repair. Although long period treatments can extend remission, a definitive cure is not yet available. One of the most common treatments is associated with dietary intake. In particular, exclusive enteral nutrition (EEN) is a first-line treatment in children with active Crohn's disease. The way EEN acts in suppressing mucosal inflammation is not fully understood, but scientists agree that modulation of intestinal microflora activity is a possible explanation [147]. To explain the influence of the diet in modulating the microbiome composition of a subject affected by Crohn's disease, in the first study I characterized the microbiome profile of an affected 14-year-old child. The same patient was enrolled before and after a nutritional therapy and the analysis was performed also on the microbiome of a healthy child with same age and sex of the affected patient. The strength of this study relies in the samples origin. In fact, the majority of the studies are based on the metagenomics profile deriving from fecal samples. In this case, I opted to use ileum samples obtained with gastrointestinal endoscopy. The reason is mainly related to the origin of Crohn's disease symptoms, which is mostly associated with the initial tract of the gut and because the microbiome composition dramatically changes across different tracts of the human gut [148], [149]. I used next generation sequencing approach, specifically 454 FLX+ Titanium (Roche), to obtain thousands of 16S rRNA sequences in order to profile the microbiome composition of the enrolled subjects.

The sequences obtained (a total of 40.621 sequences, for 3 samples) were sufficient to perform a satisfactory bioinformatics analysis.

For the analysis workflow, I chose to use the QIIME package [101], since it is the leading tool for metagenomics analysis, with more than 1.225 citations since 2010. Compared to other tools, QIIME offers the possibility to perform a complete analysis, including a wide range of sub-tools. It requires the user to have strong computational skills, but it offers to start the analysis from raw data and obtain high quality publishable results. According to other studies I found in my results, the ileum microbiome of the Crohn's disease samples is characterized by four main phylogenetic levels: *Bacteriodetes, Firmicutes, Fusobacteria, Proteobacteria* [76], [148], [150].

Specifically, the bacteria phylum level composition showed that *Proteobacteria* were more abundant and *Bacteroidetes* less abundant in the Crohn's disease patient before therapy (Patient-BT) than in the control. The most interesting thing was found in the patient after nutritional therapy (Patient-AT), (Figure 3). In fact, the composition of the ileum microbiome in the patients-AT was virtually the same as in the control. The *Fusobacteria* phylum was present only in the control subject in accordance with previous studies [151], [152]. A deeper analysis at the family level classification showed other interesting differences between the 3 subjects.

The *Bacteroidacae* family, belonging to the *Bacteriodetes* phylum, was dramatically low in the patient before therapy compared to control and patient after therapy (Figure 4). Although no statistical significance was found (due to the limited number of samples), this result has been observed in fecal samples of Cronh's disease affected patients in a previous study, confirming that alterations in the *Bacteroidacae* family may be associated with Cronh's disease [153]. In opposite, the patient before nutritional therapy reported the highest number of all the taxa in the Crohn's study belongs to the *Enterobacteriacae* family (*Proteobacteria* phylum), which alterations has also been observed in fecal samples of Cronh's disease patients [154].

Furthermore, in both control and patient after therapy the *Enterobacteriacae* family was 5 times less represented (Figure 4).

The alpha diversity analysis, as it pertains to the within-sample diversity, was computed with different methods, considering the number of sequences at a specific rarefaction depth (10.000 sequences, in accordance to the average of sequences per sample obtained with the NGS experiment).

I used both non-phylogenetic (Observed species, Chao1, Shannon Index) and phylogenetic (Faith's Phylogenetic Diversity) methods in order to highlight the bacterial community richness within every sample. The data support the reduced bacterial diversity (confirmed with all the four different methods) for the Crohn's disease patient before therapy than after therapy and even when compared with the control subject. In every result was observed a significant bacterial diversity reduction in the Crohn's disease patient before therapy ($p < 0.005$). The observed species metric (Figure 9) at a sequence depth of 10.000 sequences/sample, reported a highest number of different 16S rRNA bacteria species in the control subject, medium species diversity in the Crohn's disease patient after nutritional therapy (Patient –AT) and the lowest bacteria diversity in the patient before nutritional therapy (Patient-BT). Interestingly, the diversity of the microflora in Crohn's disease patients compared with healthy controls subjects was already observed to be 50% reduced in a previous study [151]. Our results are supported by other studies performed with different techniques besides next generation sequencing technologies. Specifically, studies with quantitative real time analysis showed a significant alteration of two major groups of anaerobic bacteria.

Bacteroides (belong to *Bacteroidacae* family), normally present in the intestinal microflora, were observed to be highly decreased in Crohn's disease and IBD affected patients [155], [156]. Furthermore, single-strand chain polymorphism (SSCP) techniques in studies of Crohn's disease microbiome, showed a close phylogenetic relationship to the *Enterobacteriacae* group,

which we found to be 5 times more represented in the patient before therapy. Generally, most of the species in this group are associated with the normal intestinal microflora, such as *Escherichia coli* and *Enterobacter*. However, other members of this group are pathogens and could drive to inflammation, such as *Shigella*, *Salmonella*, and *Yersinia* species.
Concluding, some strains of *Escherichia coli* have been suspected to play a role in the etiology of IBD and Cronh's disease [157].

7.2 Celiac disease may be associated with alterations in the gut microbiome

Celiac disease (CD) is a unique autoimmune disorder were different factors are involved. Genetic components (HLA class II genes DQ2 and/or DQ8) and environmental trigger (gluten) are known and necessary, although not sufficient for its development [158]. Other environmental components contributing to CD are thought to be present but still poorly understood. Gluten intake is known to be responsible of CD manifestation and the time of its onset, depending on both the ingested quantity and the duration of intake. In opposite to Cronh's disease, CD is not well understood. However, the role of the gut microbiota interacting with the human immune system seems to be important to maintain the mucosal integrity and functions [159]. Despite a large variability in microbiota composition across individuals, metagenomics showed that four main phyla dominate the human intestinal tract: *Bacteriodetes, Firmicutes, Fusobacteria, Proteobacteria* [76], [148], [150].
Other phyla such as *Actinobacteria, Verrucomicrobia, TM7* are only scarcely present. Compositional changes of the gut microbiota have been observed in relation to obesity and its metabolic disorders and in association with several gastrointestinal diseases [160]. In this study I analyzed the microbial composition in adult Italian celiac disease (CD) patients, both active and at a gluten-free diet (GFD-patients) and presenting the most relevant

compositional variations observed in these patients in comparison with adult controls. The analysis was performed with the same next generation sequencing approach as for the Cronh's disease study, including the bioinformatics workflow performed with the QIIME package. In the celiac disease study though, I analyzed both the bacterial (16S rRNA) and the fungal (ITS) communities. For the 16S a total of 214.999 high quality filtered sequences was obtained. The ITS sequences were more abundant with a total of 368.521 high quality filtered sequences. The quantity of sequences obtained was perfectly in the average with other studies, and allowed a deep bioinformatics analysis in order to highlight the microbiome profile of every sample. Phylum level classification among the 3 tested groups (Controls, CD-Patients, GFD-Patients) reported seven main phyla: *Actinobacteria* (9,6%), *Bacteroidetes* (16,9%), *Cyanobacteria* (0,8%), *Firmicutes* (18,6%), *Fusobacteria* (6,8%), *Proteobacteria* (45,6%), *Spirochaetes* (1%). This result agrees with other studies [148], [150]. The QIIME workflow allowed the identification of 170 different bacteria genera. As for Cronh's disease, most of the available studies regarding the role of microbiome composition in celiac disease, investigated on fecal samples by using different methodological approaches and often limiting their observations to specific bacteria. Globally, they highlighted a reduction of *Bifidobacteria* and a decreased ratio of *Lactobacilli* in CD patients respect to healthy controls [161], [162].

Globally, the data obtained on duodenal biopsies showed increased levels of *Bacteroidetes* and *Prevotella*, both belonging to the *Bacteroidetes* phylum, and decreased *Lactobacillus* belonging to *Firmicutes* phylum, whereas contradictory data have been reported concerning *Clostridium* presence [81], [163], [164]. At the state of the art, only one study used NGS methodology to investigate the intestinal microbiome in fecal samples of familial, first-degree relatives infants affected by CD with early or late gluten exposure [165].

This study is the first report on gut microbiome composition in adult CD patients, both active and at gluten-free diet and in control subjects, performed directly on duodenal biopsies using the NGS-based approach. I found that 5 main phyla contributed over 97% to the microbiome composition in all the three groups under investigation. Within these phyla, *Proteobacteria*, *Bacteroidetes* and *Fusobacteria* were increased in active CD and decreased in GFD patients, whereas *Firmicutes* where decreased in both groups respect to healthy controls. *Actinobacteria* levels were similar in all the 3 groups. The CD patients enrolled in the study showed mixed phenotypes related to gastrointestinal symptoms and/or anemia. The data obtained agreed with those observed in duodenal samples of adult CD patients with gastrointestinal symptoms regarding *Firmicutes, Fusobacteria* and *Proteobacteria* [164]. This latter Phylum (45.6%, in CD patients of this study) was respectively less and more abundant in patients with gastrointestinal symptoms (70%) or dermatitis herpetiforme (<20%) respectively, as shown in a previous study [165]. Furthermore, the same study reported *Neisseria* to be the most represented genera in active CD patients and the same result was obtained in this study. Increases in *Haemophilus* and *Neisseria* genera have been previously reported in duodenal biopsies of CD affected children [167]. *Neisseria* genus was also one of the most represented in duodenal biopsies of adult CD patients with gastrointestinal symptoms, suggesting that its interaction with the intestinal epithelium may be related with CD [166].

The alpha diversity analysis performed at a depth of 420 sequences/sample, showed an equal trend in the number of observed species in all the 3 groups.

Also the other methods used, did not show any statistical difference in the alpha diversity between the 3 groups a part a lower richness, identified in GDF patients compared to the other two groups. This lower diversity could be caused by a lower antigenic stimulation in the diet components of GFD patients. No comparison could be done with previous works, since currently

there is no study reporting the microbial diversity of ileum sample in celiac disease patients. Concerning the beta diversity, the weighted Unifrac method showed the duodenal microbiome composition of GFD patients clustering separately from CD active patients, while the control subjects reported a random distribution, suggesting that the CD condition may present more similar bacteria profiles compared to the healthy condition even after a gluten-free diet. On the other hand, the unweighted Unifrac method, showed a total random distribution of the samples, suggesting that the number of sequences/sample is an important factor in associating microbiome profiles in celiac disease. In fact, the differences found with the two Unifrac methods, might be in part related to the abundance information (i.e. the number of sequences per specific taxa) that can obscure significant patterns of variation in which taxa are present [127].

A role of fungal infections in CD pathogenesis has been also suggested [168], however there are not systemic studies investigating the fungal composition of duodenal associated microbiome in CD patients. In this study, 2 main phyla (*Ascomycota* and *Basidomycota*) were identified with no significant differences between the three study groups.

However, I identified a trend in increasing levels of the families *Mycosphaerellaceae* (within the *Ascomycota* phylum) in active-CD and GFD patients respect to the controls, even if at not statistically significant level. Using a filter of 200 seq/sample a total of 46 genera were identified in the two phyla *Ascomycota* and *Basidiomycota* (Table 10).

The most represented genera were *Cladosporiumin* and *Candida* in the *Ascomycota* and *Cryptococcus* in the *Basidiomycota*. *Candida* genus was more abundant and *Cryptococcus* less abundant in CD-Patients than in the other two groups, even if at not statistically significant level. Interestingly, CD and *Candida* may be associated. In fact, it has been demonstrated that the cell walls of *Candida*, generally responsible for oral thrush, vaginal infections and

intestinal Candidiasis, contain the same protein sequence as wheat gluten and may trigger or stimulate celiac Disease. In fact, the actual sequence of proteins that triggers celiac disease is identical to sequence of proteins, which are present in the cell walls of Candida albicans [169].

The fungal community richness was lower than the one found in the bacterial dataset, at a depth of 2.178 sequences/sample. For both the method used (Chao1 and observed species), the Controls and GFD-Patients rarefaction curves showed a very close trend while the curves obtained for the CD-Patients group presented a lower trend. The differences found were statistically significant for both the estimators comparing the alpha diversity of CD-Patients versus the other two groups ($p < 0.05$). No differences were found in the alpha diversity comparing Controls and GFD-Patients ($p = 1$), suggesting that there is lower fungal community richness in the CD-Patients group than Controls and GFD-Patients who reported quite close community richness.

The fungal beta diversity computed with Bray-Curtis method did not show any cluster among the three groups, suggesting that there are no strong similarities in the fungal microbiome profile related to the type of condition (controls, active-CD, GFD-patients).

CHAPTER 8

CONCLUSIONS

In the past few years, scientific research has contributed to highlight our understanding of the role of microorganisms inhabiting human body in health and disease conditions. Alteration in the balanced relationship between host and the microbiome can lead to an uncontrolled inflammation. Not surprisingly, the incidence of intestinal diseases has rapidly increased over the past few decades, primary due to alterations in microbial environment. The gut microbiome is fundamental to the maintenance of health, the development of disease and human metabolic processes. However, a large variation has been observed in the microbial profile of the distal gut across individuals and populations, and its composition is deeply influenced by dietary, age, sex, geographical and pharmaceutical factors. In the near future, next-generation sequencing technologies and new bioinformatics approach will most likely lead to the development of experimental models that can easily associate the human gut microbiome with onset of diseases. Future metagenomics studies need to focus on the totality of microorganisms of the human gut microbiome including fungi, viruses, yeasts and parasites in order to understand how complex ecosystems inhabiting our bodies are associated with healthy or pathological conditions. Certainly, metagenomics has already revolutionized microbiology allowing non cultivation-dependent assay and exploration of large-scale microbial communities.

Although it can provide information on the metabolic and functional capacity of a microbial community, being a DNA-based analysis, it fails to differentiate between expressed and non-expressed genes in a community.

This represents a limit and the actual metabolic activity can not be predicted [170], [171]. In this context, metatranscriptomics, defined as large-scale sequencing of mRNAs retrieved from natural communities, can better highlight microbial activities and how they are regulated [172]. Most recently, metaproteomics, the proteomic analysis of mixed microbial communities, represents a new emerging research area, which aims at assessing the immediate catalytic potential of a microbial community [173].

Only a limited amount of studies is currently available involving metatranscriptomics and metaproteomics, and new computational models are necessary in order to handle the millions of sequences continuously generated.

Eventually the future of these fields of science will be represented by the aggregation of multiple meta-approaches. In fact, meta-analysis studies, which are only starting to be performed, will be the base to understand complex biological systems and their host interactions, with the possibility to design specific drugs and to increase the power of diagnosis in a fast and automated way.

The microbiome profile defined in the Crohn's disease study represents only one step in the functional investigation of the Crohn's disease before and after a specific nutritional therapy. At the state of the art, only a limited amount of studies have analyzed the association between dietary intake and the composition of the gut microbiome in healthy subjects and in patients before and after nutritional therapy. Although my findings were obtained in one case–control study, and therefore may be considered preliminary, they strongly suggest that nutritional therapy can improve the inflammatory status of Crohn's disease by restoring the composition of the mucosal microbiome.

This case of Crohn's disease gut microbiome dysbiosis that responded to nutritional therapy can be considered proof-of-concept to evaluate a similar approach in other pediatric clinical laboratories and may serve to prompt

multicenter studies and possible clinical trials. The study has been published in the American Journal of Gastroenterology [137].

On the other hand, celiac disease available studies, investigated only fecal samples by using different methodological approaches and often limiting their observations to specific bacteria. The study I proposed is the first report on gut microbiome composition in adult CD patients (both active and at gluten free diet) and in control subjects. Furthermore, this is the first study performed directly on duodenal biopsies by using next generation sequencing based approach.

In conclusion, I decided to use the QIIME package instead of any other available tools since in a previous study I highlighted how QIIME is currently the most suitable tool for metagenomics bioinformatics analysis [174].

BIBLIOGRAPHY

[1] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., ... & Grafham, D. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860-921.

[2] Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature biotechnology*, *26*(10), 1135-1145.

[3] Asprer, J. (2012). An excitingly predictable'omic future. *Development*, *139*(20), 3675-3676.

[4] Horgan, R. P., & Kenny, L. C. (2011). 'Omic'technologies: genomics, transcriptomics, proteomics and metabolomics. *The Obstetrician & Gynaecologist*, *13*(3), 189-195.

[5] Schuster, S. C. (2007). Next-generation sequencing transforms today's biology. *Nature*, *200*(8).

[6] Niedringhaus, T. P., Milanova, D., Kerby, M. B., Snyder, M. P., & Barron, A. E. (2011). Landscape of next-generation sequencing technologies. *Analytical chemistry*, *83*(12), 4327-4341.

[7] Mardis, E. R. (2013). Next-generation sequencing platforms. *Annual review of analytical chemistry*, *6*, 287-303.

[8] Metzker, M. L. (2010). Sequencing technologies—the next generation. *Nature Reviews Genetics*, *11*(1), 31-46.

[9] Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., ... & Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics*, *13*(1), 341.

[10] Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., ... & Law, M. (2012). Comparison of next-generation sequencing systems. BioMed Research International, 2012.

[11] Glenn, T. C. (2011). Field guide to next‐generation DNA sequencers. *Molecular Ecology Resources*, *11*(5), 759-769

[12] Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews*, *68*(4), 669-685.

[13] Chen, K., & Pachter, L. (2005). Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS computational biology*, *1*(2), e24.

[14] Hugenholtz, P., Goebel, B. M., & Pace, N. R. (1998). Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *Journal of bacteriology*, *180*(18), 4765-4774.

[15] Handelsman, J., Tiedje, J. M., Alvarez-Cohen, L. I. S. A., Ashburner, M. I. C. H. A. E. L., Cann, I. K., DeLong, E. F., ... & Reid, A. H. (2007). Committee on metagenomics: challenges and functional applications. In *National Academy of Sciences, Washington* (pp. 1-158).

[16] Lederberg, J., & Mccray, A. (2001). The Scientist:\'Ome Sweet\'Omics--A Genealogical Treasury of Words. *The Scientist*, *17*(7).

[17] Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J. A., ... & Guyer, M. (2009). The NIH human microbiome project. *Genome research*,*19*(12), 2317-2323.

[18] Harmon, K. (2010). Genetics in the Gut. *Scientific American*, *302*(5), 22-24.

[19] Savage, D. C. (1977). Microbial ecology of the gastrointestinal tract. *Annual Reviews in Microbiology*, *31*(1), 107-133.

[20] Berg, R. D. (1996). The indigenous gastrointestinal microflora. *Trends in microbiology*, *4*(11), 430-435.

[21] Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C., Knight, R., & Gordon, J. I. (2007). The human microbiome project: exploring the microbial part of ourselves in a changing world. *Nature*, *449*(7164), 804.

[22] Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., ... & Lee, C. (2004). Detection of large-scale variation in the human genome.*Nature genetics*, *36*(9), 949-951.

[23] Lee, Y. K., & Mazmanian, S. K. (2010). Has the microbiota played a critical role in the evolution of the adaptive immune system?. *Science*, *330*(6012), 1768-1773.

[24] Zhu, B., Wang, X., & Li, L. (2010). Human gut microbiome: the second genome of human body. *Protein & cell*, *1*(8), 718-725.

[25] Sokol, H., Seksik, P., Rigottier-Gois, L., Lay, C., Lepage, P., Podglajen, I., ... & Doré, J. (2006). Specificities of the fecal microbiota in inflammatory bowel disease. *Inflammatory bowel diseases*, *12*(2), 106-111.

[26] Tringe, S. G., & Rubin, E. M. (2005). Metagenomics: DNA sequencing of environmental samples. *Nature reviews genetics*, *6*(11), 805-814.

[27] Kranich, J., Maslowski, K. M., & Mackay, C. R. (2011, April). Commensal flora and the regulation of inflammatory and autoimmune responses. In *Seminars in immunology* (Vol. 23, No. 2, pp. 139-145). Academic Press.

[28] O'Hara, A. M., & Shanahan, F. (2006). The gut flora as a forgotten organ.*EMBO reports*, *7*(7), 688-693.

[29] Fujimura, K. E., Slusher, N. A., Cabana, M. D., & Lynch, S. V. (2010). Role of the gut microbiota in defining human health. *Expert review of anti-infective therapy*, *8*(4), 435-454.

[30] Kau, A. L., Ahern, P. P., Griffin, N. W., Goodman, A. L., & Gordon, J. I. (2011). Human nutrition, the gut microbiome and the immune system. *Nature*,*474*(7351), 327-336.

[31] Guarner, F., & Malagelada, J. R. (2003). Gut flora in health and disease. The Lancet, 361(9356), 512-519.

[32] Steinhoff, U. (2005). Who controls the crowd? New findings and old questions about the intestinal microflora. Immunology letters, 99(1), 12-16.

[33] Shanahan, F. (2002). The host–microbe interface within the gut. *Best practice & research Clinical gastroenterology*, *16*(6), 915-931.

[34] Björkstén, B., Sepp, E., Julge, K., Voor, T., & Mikelsaar, M. (2001). Allergy development and the intestinal microflora during the first year of life. *Journal of allergy and clinical immunology*, *108*(4), 516-520.

[35] Wynne, A. G., McCartney, A. L., Brostoff, J., Hudspith, B. N., & Gibson, G. R. (2004). An in vitro assessment of the effects of broad-spectrum antibiotics on the human gut microflora and concomitant isolation of a Lactobacillus plantarum with anti-Candida activities. *Anaerobe*, *10*(3), 165-169.

[36] Hugot, J. P. (2004). Inflammatory bowel disease: a complex group of genetic disorders. *Best Practice & Research Clinical Gastroenterology*, *18*(3), 451-462.

[37] Jewell, A. P. (2005). Is the liver an important site for the development of immune tolerance to tumours?. *Medical hypotheses*, *64*(4), 751-754.

[38] Zhang, H., DiBaise, J. K., Zuccolo, A., Kudrna, D., Braidotti, M., Yu, Y., ... & Krajmalnik-Brown, R. (2009). Human gut microbiota in obesity and after gastric bypass. *Proceedings of the National Academy of Sciences*, *106*(7), 2365-2370.

[39] Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., & Gordon, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, *444*(7122), 1027-131.

[40] Mayer, E. A., Naliboff, B., & Munakata, J. (2000). The evolving neurobiology of gut feelings. *Progress in brain research*, *122*, 195-208.

[41] Suenaert, P., Bulteel, V., Lemmens, L., Noman, M., Geypens, B., Van Assche, G., ... & Rutgeerts, P. (2002). Anti-tumor necrosis factor treatment restores the gut barrier in Crohn's disease. *The American journal of gastroenterology*, *97*(8), 2000-2004.

[42] Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., ... & Gordon, J. I. (2012). Human gut microbiome viewed across age and geography. *Nature*, *486*(7402), 222-227.

[43] Cordell, B., & McCarthy, J. (2013). A Case Study of Gut Fermentation Syndrome (Auto-Brewery) with Saccharomyces cerevisiae as the Causative Organism. *International Journal of Clinical Medicine*, *4*(7).

[44] Ridaura, V. K., Faith, J. J., Rey, F. E., Cheng, J., Duncan, A. E., Kau, A. L., ... & Gordon, J. I. (2013). Gut microbiota from twins discordant for obesity modulate metabolism in mice. *Science*, *341*(6150), 1241214.

[45] Manichanh, C., Reeder, J., Gibert, P., Varela, E., Llopis, M., Antolin, M., ... & Guarner, F. (2010). Reshaping the gut microbiome with bacterial transplantation and antibiotic intake. *Genome research*, *20*(10), 1411-1419.

[46] van der Windt, D. A., Jellema, P., Mulder, C. J., Kneepkens, C. F., & van der Horst, H. E. (2010). Diagnostic testing for celiac disease among patients with abdominal symptoms: a systematic review. *Jama*, *303*(17), 1738-1746.

[47] Rewers, M. J. Epidemiology of Celiac Disease: What Are the Prevalence, Incidence, and Progression of Celiac Disease?. *National Institutes of Health*, 45.

[48] Houlston, R. S., & Ford, D. (1996). Genetics of coeliac disease. *Qjm*, *89*(10), 737-744.

[49] Bamford, F. N. (1989). Child and Adolescent Health for Practitioners. *Archives of disease in childhood*, *64*(2), 312.

[50] Van Heel, D. A., & West, J. (2006). Recent advances in coeliac disease. *Gut*,*55*(7), 1037-1046.

[51] Hadithi, M., Von Blomberg, B. M. E., Crusius, J. B. A., Bloemena, E., Kostense, P. J., Meijer, J. W., ... & Stehouwer, C. D. (2007). Accuracy of serologic tests and HLA-DQ typing for diagnosing celiac disease. *Annals of internal medicine*, *147*(5), 294-302.

[52] Kupper, C. (2005). Dietary guidelines and implementation for celiac disease.*Gastroenterology*, *128*(4), S121-S127.

[53] Häuser, W., Stallmach, A., Caspary, W. F., & Stein, J. (2007). Predictors of reduced health-related quality of life in adults with coeliac disease. *Alimentary pharmacology & therapeutics*, *25*(5), 569-578.

[54] Baumgart, D. C., & Sandborn, W. J. (2012). Crohn's disease. *The Lancet*, *380*(9853), 1590-1605.

[55] Calkins, B. M., Lilienfeld, A. M., Garland, C. F., & Mendeloff, A. I. (1984). Trends in incidence rates of ulcerative colitis and Crohn's disease. *Digestive diseases and sciences*, *29*(10), 913-920.

[56] Henckaerts, L., Figueroa, C., Vermeire, S., & Sans, M. (2008). The role of genetics in inflammatory bowel disease. *Current drug targets*, *9*(5), 361-368.

[57] Barrett, J. C., Hansoul, S., Nicolae, D. L., Cho, J. H., Duerr, R. H., Rioux, J. D., ... & Ghori, J. (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature genetics*, *40*(8), 955-962.

[58] Baumgart, D. C., & Carding, S. R. (2007). Inflammatory bowel disease: cause and immunobiology. *The Lancet*, *369*(9573), 1627-1640.

[59] Xavier, R. J., & Podolsky, D. K. (2007). Unravelling the pathogenesis of inflammatory bowel disease. *Nature*, *448*(7152), 427-434.

[60] O'Sullivan, M., & O'Morain, C. (2006). Nutrition in inflammatory bowel disease.*Best Practice & Research Clinical Gastroenterology*, *20*(3), 561-573.

[61] Mukhopadhya, I., Hansen, R., El-Omar, E. M., & Hold, G. L. (2012). IBD—what role do Proteobacteria play?. *Nature Reviews Gastroenterology and Hepatology*, *9*(4), 219-230.

[62] Jostins, L., Ripke, S., Weersma, R. K., Duerr, R. H., McGovern, D. P., Hui, K. Y., ... & Hart, A. (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, *491*(7422), 119-124.

[63] Loftus, E. V., Silverstein, M. D., Sandborn, W. J., Tremaine, W. J., Harmsen, W. S., & Zinsmeister, A. R. (2000). Ulcerative colitis in Olmsted County, Minnesota, 1940–1993: incidence, prevalence, and survival. *Gut*, *46*(3), 336-343.

[64] Barrett, J. C., Hansoul, S., Nicolae, D. L., Cho, J. H., Duerr, R. H., Rioux, J. D., ... & Ghori, J. (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature genetics*, *40*(8), 955-962.

[65] Wang, K., Li, M., & Hakonarson, H. (2010). Analysing biological pathways in genome-wide association studies. *Nature Reviews Genetics*, *11*(12), 843-854.

[66] Ogura, Y., Bonen, D. K., Inohara, N., Nicolae, D. L., Chen, F. F., Ramos, R., ... & Cho, J. H. (2001). A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature*, *411*(6837), 603-606.

[67] Hanauer, S. B. (2006). Inflammatory bowel disease: epidemiology, pathogenesis, and therapeutic opportunities. *Inflammatory bowel diseases*,*12*(5), S3-S9.

[68] Dogan, B., Scherl, E., Bosworth, B., Yantiss, R., Altier, C., McDonough, P. L., ... & Simpson, K. W. (2012). Multidrug resistance is common in Escherichia coli associated with ileal Crohn's disease. *Inflammatory Bowel Diseases*.

[69] Weinstock, G. M. (2012). Genomic approaches to studying the human microbiota. *Nature*, *489*(7415), 250-256.

[70] Goodman, A. L., Kallstrom, G., Faith, J. J., Reyes, A., Moore, A., Dantas, G., & Gordon, J. I. (2011). Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice. *Proceedings of the National Academy of Sciences*, *108*(15), 6252-6257.

[71] Abubucker, S., Segata, N., Goll, J., Schubert, A. M., Izard, J., Cantarel, B. L., ... & Huttenhower, C. (2012). Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS computational biology*, *8*(6), e1002358.

[72] Burke, C., Steinberg, P., Rusch, D., Kjelleberg, S., & Thomas, T. (2011). Bacterial community assembly based on functional genes rather than species.*Proceedings of the National Academy of Sciences*, *108*(34), 14288-14293.

[73] Eckburg, P. B., Bik, E. M., Bernstein, C. N., Purdom, E., Dethlefsen, L., Sargent, M., ... & Relman, D. A. (2005). Diversity of the human intestinal microbial flora. *Science*, *308*(5728), 1635-1638.

[74] Human Microbiome Project Consortium. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, *486*(7402), 207-214.

[75] Frank, D. N., Amand, A. L. S., Feldman, R. A., Boedeker, E. C., Harpaz, N., & Pace, N. R. (2007). Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proceedings of the National Academy of Sciences*, *104*(34), 13780-13785.

[76] Manichanh, C., Rigottier-Gois, L., Bonnaud, E., Gloux, K., Pelletier, E., Frangeul, L., ... & Dore, J. (2006). Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut*, *55*(2), 205-211.

[77] MacLean, D., Jones, J. D., & Studholme, D. J. (2009). Application of'next-generation'sequencing technologies to microbial genetics. *Nature Reviews Microbiology*, *7*(4), 287-296.

[78] Hawrelak, J. A., & Myers, S. P. (2004). The causes of intestinal dysbiosis: a review. *Alternative medicine review*, *9*(2).

[79] Forsberg, G., Fahlgren, A., Hörstedt, P., Hammarström, S., Hernell, O., & Hammarström, M. L. (2004). Presence of bacteria and innate immunity of intestinal epithelium in childhood celiac disease. *The American journal of gastroenterology*, *99*(5), 894-904.

[80] Collado, M. C., Donat, E., Ribes-Koninckx, C., Calabuig, M., & Sanz, Y. (2009). Specific duodenal and faecal bacterial groups associated with paediatric coeliac disease. *Journal of clinical pathology*, *62*(3), 264-269.

[81] Nadal, I., Donant, E., Ribes-Koninckx, C., Calabuig, M., & Sanz, Y. (2007). Imbalance in the composition of the duodenal microbiota of children with coeliac disease. *Journal of Medical Microbiology*, *56*(12), 1669-1674.

[82] Sanz, Y., Sánchez, E., Marzotto, M., Calabuig, M., Torriani, S., & Dellaglio, F. (2007). Differences in faecal bacterial communities in coeliac and healthy children as detected by PCR and denaturing gradient gel electrophoresis.*FEMS Immunology & Medical Microbiology*, *51*(3), 562-568.

[83] Bertini, I., Calabro, A., De Carli, V., Luchinat, C., Nepi, S., Porfirio, B., ... & Tenori, L. (2008). The metabonomic signature of celiac disease. *Journal of proteome research*, *8*(1), 170-177.

[84] McKenna, P., Hoffmann, C., Minkah, N., Aye, P. P., Lackner, A., Liu, Z., ... & Bushman, F. D. (2008). The macaque gut microbiome in health, lentiviral infection, and chronic enterocolitis. *PLoS pathogens*, *4*(2), e20.

[85] Di Cagno, R., De Angelis, M., De Pasquale, I., Ndagijimana, M., Vernocchi, P., Ricciuti, P., ... & Francavilla, R. (2011). Duodenal and faecal microbiota of celiac children: molecular, phenotype and metabolome characterization. *BMC microbiology*, *11*(1), 219.

[86] Collado, M. C., Donat, E., Ribes-Koninckx, C., Calabuig, M., & Sanz, Y. (2009). Specific duodenal and faecal bacterial groups associated with paediatric coeliac disease. *Journal of clinical pathology*, *62*(3), 264-269.

[87] Béjà, O., Suzuki, M. T., Koonin, E. V., Aravind, L., Hadd, A., Nguyen, L. P., ... & DeLong, E. F. (2000). Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environmental Microbiology*, *2*(5), 516-529.

[88] Poinar, H. N., Schwarz, C., Qi, J., Shapiro, B., MacPhee, R. D., Buigues, B., ... & Schuster, S. C. (2006). Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *science*, *311*(5759), 392-394.

[89] Wooley, J. C., Godzik, A., & Friedberg, I. (2010). A primer on metagenomics.*PLoS computational biology*, *6*(2), e1000667.

[90] Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., ... & Weissenbach, J. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, *464*(7285), 59-65.

[91] Haas, B. J., Gevers, D., Earl, A. M., Feldgarden, M., Ward, D. V., Giannoukos, G., ... & Birren, B. W. (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome research*, *21*(3), 494-504.

[92] Soergel, D. A., Dey, N., Knight, R., & Brenner, S. E. (2012). Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. *The ISME journal*, *6*(7), 1440-1444.

[93] Werner, J. J., Zhou, D., Caporaso, J. G., Knight, R., & Angenent, L. T. (2012). Comparison of Illumina paired-end and single-direction sequencing for microbial 16S rRNA gene amplicon surveys. *The ISME journal*, *6*(7), 1273-1276.

[94] Asai, T., Zaporojets, D., Squires, C., & Squires, C. L. (1999). An Escherichia coli strain with all chromosomal rRNA operons inactivated: complete exchange of rRNA genes between bacteria. *Proceedings of the National Academy of Sciences*, *96*(5), 1971-1976.

[95] Schouls, L. M., Schot, C. S., & Jacobs, J. A. (2003). Horizontal transfer of segments of the 16S rRNA genes between species of the Streptococcus anginosus group. *Journal of bacteriology*, *185*(24), 7241-7246.

[96] Fox, G. E., Wisotzkey, J. D., & Jurtshuk, P. (1992). How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity.*International Journal of Systematic Bacteriology*, *42*(1), 166-170.

[97] Hallam, S. J., Putnam, N., Preston, C. M., Detter, J. C., Rokhsar, D., Richardson, P. M., & DeLong, E. F. (2004). Reverse methanogenesis: testing the hypothesis with environmental genomics. Science, 305(5689), 1457-1462.

[98] Bray, N., & Pachter, L. (2004). MAVID: constrained ancestral alignment of multiple sequences. *Genome research*, *14*(4), 693-699.

[99] Sundararajan, M., Brudno, M., Small, K., Sidow, A., & Batzoglou, S. (2004). Chaining algorithms for alignment of draft sequence. In *Algorithms in Bioinformatics* (pp. 326-337). Springer Berlin Heidelberg.

[100] Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., ... & Zagnitko, O. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC genomics*, *9*(1), 75.

[101] Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., ... & Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature methods*, *7*(5), 335-336.

[102] Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., ... & Edwards, R. A. (2008). The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC bioinformatics*, *9*(1), 386.

[103] Huson, D. H., Mitra, S., Ruscheweyh, H. J., Weber, N., & Schuster, S. C. (2011). Integrative analysis of environmental sequences using MEGAN4.*Genome research*, *21*(9), 1552-1560.

[104] Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., ... & Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology*, *75*(23), 7537-7541.

[105] Ewing, B., Hillier, L., Wendl, M. C., & Green, P. (1998). Base-calling of automated sequencer traces usingPhred. I. Accuracy assessment. *Genome research*, *8*(3), 175-185.

[106] Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., Canese, K., ... & Ye, J. (2010). Database resources of the national center for biotechnology information. *Nucleic acids research*, *38*(suppl 1), D5-D16.

[107] Li, W., & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658-1659.

[108] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, *215*(3), 403-410.

[109] Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., ... & Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology*, *75*(23), 7537-7541.

[110] Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, *26*(19), 2460-2461.

[111] Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., & Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*,*27*(16), 2194-2200.

[112] Chen, W., Zhang, C. K., Cheng, Y., Zhang, S., & Zhao, H. (2013). A Comparison of Methods for Clustering 16S rRNA Sequences into OTUs. *PloS one*, *8*(8), e70837.

[113] Statnikov, A., Henaff, M., Narendra, V., Konganti, K., Li, Z., Yang, L., ... & Alekseyenko, A. V. (2013). A comprehensive evaluation of multicategory classification methods for microbiomic data. *Microbiome*, *1*(1), 11.

[114] Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microb 73(16): 5261-5267.

[115] Soergel DAW, Dey N, Knight R, Brenner SE. (2012). Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. ISME J (6), 1440–1444.

[116] DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., ... & Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and environmental microbiology*, *72*(7), 5069-5072.

[117] Edgar, Robert C. (2004), MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Research 32(5), 1792-97.

[118] Caporaso, J. G., Bittinger, K., Bushman, F. D., DeSantis, T. Z., Andersen, G. L., & Knight, R. (2010). PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics*, *26*(2), 266-267.

[119] McDonald, D., Clemente, J. C., Kuczynski, J., Rideout, J. R., Stombaugh, J., Wendel, D., ... & Caporaso, J. G. (2012). The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome.*GigaScience*, *1*(1), 7.

[120] Whittaker, R. H. (1972). Evolution and measurement of species diversity.*Taxon*, 213-251.

[121] Hill, M. O. (1973). Diversity and evenness: a unifying notation and its consequences. *Ecology*, *54*(2), 427-432.

[122] Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of statistics*, 265-270.

[123] Shannon, C. E. (1948). A mathematical theory of communication. The Bell System Technical Journal, 27, 379–423 and 623–656.

[124] Daniel P. Faith, Andrew M. Baker. (2007). Phylogenetic diversity (PD) and biodiversity conservation: some bioinformatics challenges. Evolutionary Bioinformatics, 2, 121-128.

[125] Hamady, M., & Knight, R. (2009). Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome research*,*19*(7), 1141-1152.

[126] Lozupone, C., & Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and environmental microbiology*,*71*(12), 8228-8235.

[127] Lozupone, C. A., & Knight, R. (2007). Global patterns in bacterial diversity.*Proceedings of the National Academy of Sciences*, *104*(27), 11436-11440.

[128] Bray, J. R., & Curtis, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecological monographs*, *27*(4), 325-349.

[129] STAHLE, L., & Wold, S. (1989). Analysis of variance (ANOVA). *Chemometrics and intelligent laboratory systems*, *6*(4), 259-272.

[130] Tyson, H. (2013). Biometry, The Principles and Practice of Statistics in Biological Research.

[131] Abdi, H. (2007). The Bonferonni and Šidák corrections for multiple comparisons. *Encyclopedia of measurement and statistics*, *3*, 103-107.

[132] Digby, P. G. N., & Kempton, R. A. (1987). *Multivariate analysis of ecological communities* (pp. 80-86). London: Chapman and Hall.

[133] Anderson, M. J. (2001). A new method for non‐parametric multivariate analysis of variance. *Austral ecology*, *26*(1), 32-46.

[134] Milton, J. S., & Arnold, J. C. (2002). *Introduction to probability and statistics: principles and applications for engineering and the computing sciences*. McGraw-Hill, Inc..

[135] Day, A. S., Whitten, K. E., Sidler, M., & Lemberg, D. A. (2008). Systematic review: nutritional therapy in paediatric Crohn's disease. *Alimentary pharmacology & therapeutics*, *27*(4), 293-307.

[136] Lionetti, P., Callegari, M. L., Ferrari, S., Cavicchi, M. C., Pozzi, E., de Martino, M., & Morelli, L. (2005). Enteral nutrition and microflora in pediatric Crohn's disease. *Journal of Parenteral and Enteral Nutrition*, *29*(4 suppl), S173-S178.

[137] D'Argenio, V., Precone, V., Casaburi, G., Miele, E., Martinelli, M., Staiano, A., ... & Sacchetti, L. (2013). An altered gut microbiome profile in a child affected by Crohn's disease normalized after nutritional therapy. *The American journal of gastroenterology*, *108*(5), 851.

[138] Baker, G. C., Smith, J. J., & Cowan, D. A. (2003). Review and re-analysis of domain-specific 16S primers. *Journal of Microbiological Methods*, *55*(3), 541-555.

[139] Valášková, V., de Boer, W., Gunnewiek, P. J. K., Pospíšek, M., & Baldrian, P. (2009). Phylogenetic composition and properties of bacteria coexisting with the fungus Hypholoma fasciculare in decaying wood. *The ISME journal*, *3*(10), 1218-1221.

[140] Sun, Y., Cai, Y., Huse, S. M., Knight, R., Farmerie, W. G., Wang, X., & Mai, V. (2012). A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. *Briefings in bioinformatics*, *13*(1), 107-121.

[141] DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., ... & Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and environmental microbiology*, *72*(7), 5069-5072.

[142] Kõljalg, U., Larsson, K. H., Abarenkov, K., Nilsson, R. H., Alexander, I. J., Eberhardt, U., ... & Vrålstad, T. (2005). UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. *New Phytologist*, *166*(3), 1063-1068.

[143] Cardona, G., Rosselló, F., & Valiente, G. (2008). Extended Newick: it is time for a standard representation of phylogenetic networks. *BMC bioinformatics*,*9*(1), 532.

[144] DeSantis, T. Z., Hugenholtz, P., Keller, K., Brodie, E. L., Larsen, N., Piceno, Y. M., ... & Andersen, G. L. (2006). NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. *Nucleic Acids Research*,*34*(suppl 2), W394-W399.

[145] Khor, B., Gardet, A., & Xavier, R. J. (2011). Genetics and pathogenesis of inflammatory bowel disease. *Nature*, *474*(7351), 307-317.

[146] Cosnes, J., Gower–Rousseau, C., Seksik, P., & Cortot, A. (2011). Epidemiology and natural history of inflammatory bowel diseases.*Gastroenterology*, *140*(6), 1785-1794.

[147] Jones, V. A., Workman, E., Freeman, A. H., Dickinson, R. J., Wilson, A. J., & Hunter, J. O. (1985). Crohn's disease: maintenance of remission by diet. *The Lancet*, *326*(8448), 177-180.

[148] Willing, B. P., Dicksved, J., Halfvarson, J., Andersson, A. F., Lucio, M., Zheng, Z., ... & Engstrand, L. (2010). A pyrosequencing study in twins shows that gastrointestinal microbial profiles vary with inflammatory bowel disease phenotypes. *Gastroenterology*, *139*(6), 1844-1854.

[149] Hansen, J., Gulati, A., & Sartor, R. B. (2010). The role of mucosal immunity and host genetics in defining intestinal commensal bacteria. *Current opinion in gastroenterology*, *26*(6), 564.

[150] Cho, I., & Blaser, M. J. (2012). The human microbiome: at the interface of health and disease. *Nature Reviews Genetics*, *13*(4), 260-270.

[151] Ott, S. J., Musfeldt, M., Wenderoth, D. F., Hampe, J., Brant, O., Fölsch, U. R., & Schreiber, S. (2004). Reduction in diversity of the colonic mucosa associated bacterial microflora in patients with active inflammatory bowel disease. *Gut*, *53*(5), 685-693.

[152] Neut, C., Bulois, P., Desreumaux, P., Membreé, J. M., Lederman, E., Gambiez, L., & Colombel, J. F. (2002). Changes in the bacterial flora of the neoterminal ileum after ileocolonic resection for Crohn's disease. *The American journal of gastroenterology*, *97*(4), 939-946.

[153] Takaishi, H., Matsuki, T., Nakazawa, A., Takada, T., Kado, S., Asahara, T., & Hibi, T. (2008). Imbalance in intestinal microflora constitution could be involved in the pathogenesis of inflammatory bowel disease. *International Journal of Medical Microbiology*, *298*(5), 463-472.

[154] Seksik, P., Rigottier-Gois, L., Gramet, G., Sutren, M., Pochart, P., Marteau, P., ... & Dore, J. (2003). Alterations of the dominant faecal bacterial groups in patients with Crohn's disease of the colon. *Gut*, *52*(2), 237-242.

[155] Zoetendal EG, von Wright A, Vilpponen-Salmela T, et al. Mucosa-associated bacteria in the human gastrointestinal tract are uniformly distributed along the colon and differ from the community recovered from feces. Appl Environ Microbiol2002;68:3401–7.

[156] Kleessen, B., Kroesen, A. J., Buhr, H. J., & Blaut, M. (2002). Mucosal and invading bacteria in patients with inflammatory bowel disease compared with controls. *Scandinavian journal of gastroenterology*, *37*(9), 1034-1041.

[157] Sartor RB. Microbial factors in the pathogenesis of Crohn's disease, ulcerative colitis, and experimental intestinal inflammation. In: Kirshner JB, ed. *Inflammatory* bowel disease, 5th edn. Philadelphia: WB Saunders, 2000:153–78.

[158] Catassi, C., & Fasano, A. (2008). Celiac disease. *Current opinion in gastroenterology*, *24*(6), 687-691.

[159] Sellitto, M., Bai, G., Serena, G., Fricke, W. F., Sturgeon, C., Gajer, P., ... & Fasano, A. (2012). Proof of concept of microbiome-metabolome analysis and delayed gluten exposure on celiac disease autoimmunity in genetically at-risk infants. *PloS one*, *7*(3), e33387.

[160] Harris, K., Kassis, A., Major, G., & Chou, C. J. (2012). Is the gut microbiota a new factor contributing to obesity and its metabolic disorders?. *Journal of obesity*, *2012*.

[161] Di Cagno, R., Rizzello, C. G., Gagliardi, F., Ricciuti, P., Ndagijimana, M., Francavilla, R., ... & De Angelis, M. (2009). Different fecal microbiotas and volatile organic compounds in treated and untreated children with celiac disease. Applied and environmental *microbiology*, 75(12), 3963-3971.

[162] De Palma, G., Cinova, J., Stepankova, R., Tuckova, L., & Sanz, Y. (2010). Pivotal Advance: Bifidobacteria and Gram-negative bacteria differentially influence immune responses in the proinflammatory milieu of celiac disease. *Journal of leukocyte biology*, *87*(5), 765-778.

[163] Shetty, S. A., Marathe, N. P., & Shouche, Y. S. (2013). Opportunities and challenges for gut microbiome studies in the Indian population. *Microbiome*, *1*(1), 24.

[164] Calabrò, A., Gralka, E., Luchinat, C., Saccenti, E., & Tenori, L. A metabolomic perspective on coeliac disease.

[165] Kupfer, S. S., & Jabri, B. (2012). Pathophysiology of celiac disease. *Gastrointestinal endoscopy clinics of North America*, *22*(4), 639-660.

[166] Wacklin, P., Kaukinen, K., Tuovinen, E., Collin, P., Lindfors, K., Partanen, J., ... & Mättö, J. (2013). The duodenal microbiota composition of adult celiac disease patients is associated with the clinical manifestation of the disease. *Inflammatory bowel diseases*, *19*(5), 934-941.

[167] Nistal, E., Caminero, A., Herrán, A. R., Arias, L., Vivas, S., De Morales, J. M. R., ... & Casqueiro, J. (2012). Differences of small intestinal bacteria populations in adults and children with/without celiac disease: effect of age, gluten diet, and disease. *Inflammatory bowel diseases*, *18*(4), 649-656.

[168] Rizzello, C. G., De Angelis, M., Di Cagno, R., Camarca, A., Silano, M., Losito, I., ... & Gobbetti, M. (2007). Highly efficient gluten degradation by lactobacilli and fungal proteases during food processing: new perspectives for celiac disease. *Applied and Environmental Microbiology*, *73*(14), 4499-4507.

[169] Nieuwenhuizen, W. F., Pieters, R. H. H., Knippels, L. M. J., Jansen, M. C. J. F., & Koppelman, S. J. (2003). Is Candida albicans a trigger in the onset of coeliac disease?. *The Lancet*, *361*(9375), 2152-2154.

[170] Simon, C., & Daniel, R. (2011). Metagenomic analyses: past and future trends. *Applied and environmental microbiology*, *77*(4), 1153-1161.

[171] Sorek, R., & Cossart, P. (2009). Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nature Reviews Genetics*, *11*(1), 9-16.

[172] Bailly, J., Fraissinet-Tachet, L., Verner, M. C., Debaud, J. C., Lemaire, M., Wésolowski-Louvel, M., & Marmeisse, R. (2007). Soil eukaryotic functional diversity, a metatranscriptomic approach. *The ISME journal*, *1*(7), 632-642.

[173] Wilmes, P., & Bond, P. L. (2004). The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. *Environmental Microbiology*, *6*(9), 911-920.

[174] D'Argenio, V., Casaburi, G., Precone, V., Salvatore, F., Federico, I. I., & Salvatore, F. (2014) Comparative Metagenomic Analysis of Human Gut Microbiome Composition Using Two Different Bioinformatic Pipelines. *BioMed Research International*.