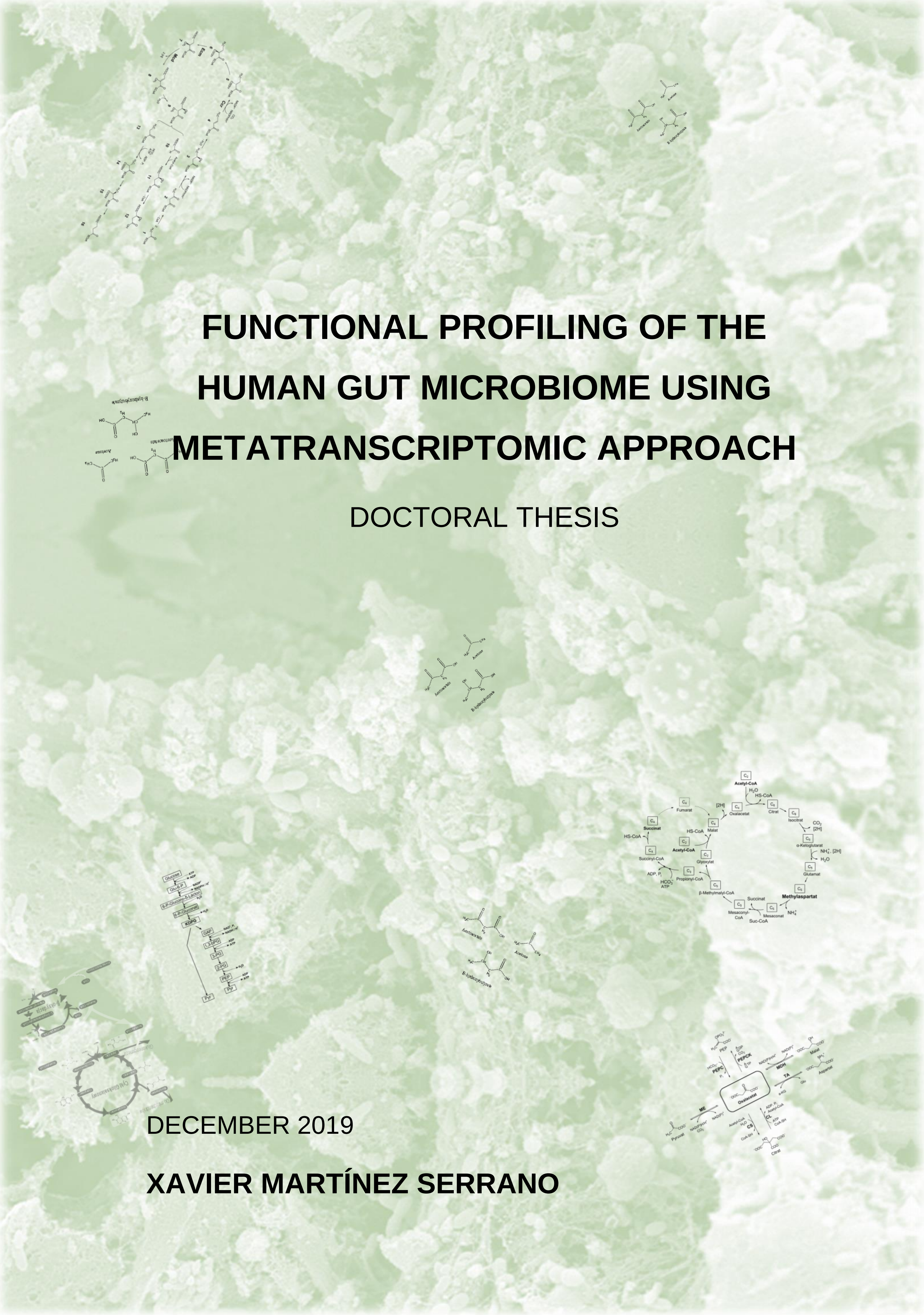
FUNCTIONAL PROFILING OF THE HUMAN GUT MICROBIOME USING METATRANSCRIPTOMIC APPROACH

XAVIER MARTÍNEZ SERRANO



FUNCTIONAL PROFILING OF THE HUMAN GUT MICROBIOME USING METATRANSCRIPTOMIC APPROACH

DOCTORAL THESIS

DECEMBER 2019

XAVIER MARTÍNEZ SERRANO

PROGRAMA DE DOCTORAT EN MEDICINA

DEPARTAMENT DE MEDICINA

FUNCTIONAL PROFILING OF THE HUMAN GUT MICROBIOME USING METATRANSCRIPTOMIC APPROACH

XAVIER MARTÍNEZ SERRANO

PhD THESIS – Barcelona, December 2019

Director	Tutor	Author
Dr. Chaysavanh Manichanh	**Dr. Victor Manuel Vargas Blasco**	**Xavier Martínez Serrano**

Universitat Autònoma de Barcelona

Dr. Chaysavanh Manichanh

Principal Investigator – Microbiome Lab
Department of Physiology and Physiopathology of the Digestive Tract
Vall d’Hebron Institut de Recerca

Certify that the thesis entitled **“Functional profiling of the human gut microbiome using metatranscriptomic approach”** submitted by **Xavier Martínez Serrano**, was carried out under her supervision and tutorship of **Dr. Víctor Vargas Blasco**, and authorize its presentation for the defense ahead of the corresponding tribunal.

Barcelona, December 2019

“All sorts of computer errors are now turning up.

You'd be surprised to know the number of doctors

who claim they are treating pregnant men”

I. Asimov

“Science knows no country,

because knowledge belongs to humanity,

and is the torch which illuminates the world.”

L. Pasteur

AGRAÏMENTS

La finalització d'aquesta tesi ha estat possible gràcies a la col·laboració de moltes persones que també hi han contribuït amb l'únic propòsit que aquest treball pogués arribar a bon port.

Un especial agraïment a la Laia, la meva parella i mare dels meus dos fills, que ha obert l'espai necessari per poder encabir tot el temps addicional que necessitava i que la meva responsabilitat com a pare no em permetia.

Als meus dos fills, que també han sofert la meva falta de temps invertit que els hi pertocava.

A la inestimable ajuda i suport dels meus pares i sogres que sempre han estat al nostre costat fent la pinya per pujar el castell.

Un agraïment especial a la Dra. Chaysavanh per la seva paciència i atenció en resoldre els meus dubtes i solucionar els problemes que hem anat trobant.

A totes les companyes i companys de feina del MetaLab que també han contribuït a donar-me suport i compartir els seus coneixements. I més important encara, agrair-vos totes les bones estones que hem pogut compartir durant el dia a dia.

Aquesta tesi està especialment dedicada a dues persones estimades, que durant aquest temps ens han deixat de forma tan inesperada. Joana i Lawrence, us portarem sempre amb nosaltres.

ABBREVIATIONS

AU, Approximately Unbiased

AUC, Area Under the Curve

AWK, Aho, Weinberger and Kernighan (scripting language)

BLAST, Basic Local Alignment Search Tool

BLASTP, BLAST for Protein alignment

BLAT, BLAST-like alignment tool

BMI, Body Mass Index

BP, Bootstrap Probability

CAI, Colitis Activity Index

CD, Crohn’s Disease

COG, Clusters of Orthologous Groups

DEA, differential expression analysis

eggNOG, evolutionary genealogy of genes Non-supervised Orthologous Genes

F, as part of F.REM or F.REL stands for "Future", in another context “Filter”

FBD, Functional Bowel Disorders

FDR, False Discovery Rate

FMT, Fecal Microbial Transplantation

GIT, GastroIntestinal Tract

H, HC, HR, healthy, healthy controls, healthy relatives (all are synonyms)

HMP, Human Microbiome Project

HU, HUMAnN, HMP Unified Metabolic Analysis Network

IBD, Inflammatory Bowel Disease

IBS, Irritable Bowel Syndrome

IGC, Integrated Gene Catalog

KEGG, Kyoto Encyclopedia of Genes and Genomes

LTP, Last timepoint

MetaHit14, refers to the Integrated Gene Catalog from MetaHit released in 2014

MOCAT, Metagenomic Analysis Tookit

PATRIC, Pathosystems resource integration center

PCR, Polymearse Chain Reaction

POSIX, Portable Operating System Interface

QIIME, Quantitative Insights Into Microbial Ecology

REL, relapse

REM, remission

TP0, Timepoint zero (basal or baseline)

TPM, transcripts per million

TPR, true positive rate

UC, Ulcerative Colitis

UCLUST, algorithm divides a set of sequences into clusters.

UniRef50, UniProt Reference Clusters, protein database at 50% identity

UniRef90, UniProt Reference Clusters, protein database at 90% identity

USEARCH, Ultra fast sequence analysis

TABLE OF CONTENTS

Abstract

To date, meta-omic approaches use high-throughput sequencing technologies, which produce a tremendous amount of data, thus challenging modern computers. We developed a new open-source pipeline, namely MetaTrans, to analyze the structure and functions of active microbial communities using the power of multi-threading computers. The pipeline is designed to perform two types of RNA-Seq analyses: taxonomic and gene expression. It performs quality-control assessment, rRNA removal, maps reads against functional databases and also handles differential gene expression analysis. Its efficacy was validated analyzing data from synthetic mock communities, data generated from a previous study on irritable bowel syndrome (IBS), and comparing with a recently published metagenomics and metatranscriptomics study. Compared to an existing web application server, MetaTrans shows more efficiency in terms of runtime (around 2 hours per million of transcripts) and presents adapted tools to compare gene expression levels. It has been tested with a human gut microbiome database but also proposes an option to use a general database in order to analyze other ecosystems. For the installation and use of the pipeline, we provide a detailed guide at the following website (www.metatrans.org). We next applied this pipeline to investigate the taxonomic and functional profilings of the active microbiota of patients with Crohn's disease (CD) and ulcerative colitis (UC), two main forms of inflammatory bowel disease (IBD). For this purpose, healthy controls and patients with CD and UC provided fecal samples at two time points, from which cDNA were generated and sequenced. Our analysis of the sequence data revealed that CD and UC presented a distinct active microbiome profile at the taxonomic as well as functional level. Furthermore, CD patients showed greater dysbiosis than UC patients. Our results also suggested that dysregulations of different pathways related to the Short Chain Fatty Acids metabolism and cell survival were associated with disease severity. Altogether, our study provides a very comprehensive description of the active microbial functions and paves the way for future investigations on irritable bowel syndrome and inflammatory bowel diseases.

Resum

Fins ara, les aproximacions meta-òmiques utilitzen tecnologies de seqüenciació d'alt rendiment, que produeixen una quantitat enorme de dades, desafiant així els ordinadors moderns actuals. Hem desenvolupat una nova pipeline (unió de diverses eines per realitzar una determinada tasca) de codi obert, que hem anomenat MetaTrans, per analitzar l'estructura i les funcions de les comunitats microbianes actives utilitzant la potència dels ordinadors multi-fil. La pipeline està dissenyada per realitzar dos tipus d'anàlisis de seqüenciació de RNA (RNA-Seq): el taxonòmic i el d'expressió gènica. Fa un control de qualitat, elimina l'rRNA, alinea lectures de seqüenciació (anomenades reads) contra bases de dades funcionals i també dur a terme anàlisis d'expressió gènica diferencial. La seva eficàcia va ser validada per mitjà de l'anàlisi de dades de simulacions sintètiques de comunitats, dades generades en un altre estudi de la síndrome de l'intestí irritable (SII), i comparant amb un estudi, publicat recentment, sobre metagenòmica i metatranscriptòmica. En comparació amb un servidor d'aplicacions web existent, MetaTrans mostra més eficiència en termes de temps d'execució (al voltant de 2 hores per milió de transcripcions) i presenta eines adaptades per comparar nivells d'expressió gènica. S'ha provat amb una base de dades de microbioma de l'intestí humà, però també proposa una opció per utilitzar una base de dades general per tal d'analitzar altres ecosistemes. Per a la instal·lació i l'ús de la pipeline, proporcionem una guia detallada a la següent pàgina web (www.metatrans.org). A continuació, vam utilitzar aquesta pipeline per investigar els perfils taxonòmics i funcionals de la microbiota activa en pacients amb la malaltia de Crohn (MC) i colitis ulcerosa (CU), dues formes principals de la malaltia inflamatòria intestinal (MII). Per a aquest propòsit, els controls sans i els pacients amb MC i CU van proporcionar mostres fecals en dos punts de temps, en els quals es va generar el DNA complementari (cDNA) i es va seqüenciar. La nostra anàlisi de les dades seqüenciades va revelar que MC i la CU presentaven un diferent perfil actiu de microbioma tant a nivell taxonòmic com funcional. A més a més, els pacients amb MC van mostrar una major disbiosis que els pacients de amb CU. Els nostres resultats també van suggerir que la desregulació de diferents vies o rutes metabòliques relacionades amb el metabolisme dels àcids grassos de cadena curta i la supervivència cel·lular

estava associada amb la gravetat de la malaltia. En conjunt, el nostre estudi proporciona una descripció molt completa de les funcions microbianes que són actives, i prepara el camí per a futures investigacions sobre la síndrome de l'intestí irritable i les malalties inflamatòries intestinals.

TABLE OF TABLES

TABLE OF FIGURES

Chapter 1.

Introduction

1 Introduction

1.1 Microbiota and microbiome

The definition of the term microbiota was first coined by Lederberg in 2001 (Lederberg and McCray, 2001; Prescott, 2017) and refers to the group of microorganisms present in a defined environment (Marchesi and Ravel, 2015).

The different body sites are composed of different microbial community compositions. We, as hosts, stablish a close relationship with the microbial communities in a symbiotic way since we are born. We live with them, we feed them, and we obtain beneficial nutrients and substrates for our organism. Past studies have analyzed the bacterial distribution in each body site and found the gastrointestinal (GI) tract as the organ harboring the major abundance of microbes in the body, which has become one of the most studied ecosystems (Peterson *et al.*, 2009; Bokulich *et al.*, 2013). The human microbiota encompasses prokaryotes such as bacteria and archaea, eukaryotes such as fungi, and viruses (bacteriophages).

Early studies quoted that microbes in our bodies made up 10x the number of human cells (estimated to 100 trillion) (Ley, Peterson and Gordon, 2006) and encoded up to 150x the number of human genes (Qin *et al.*, 2010). Nevertheless, recent articles showed that human:bacteria cells ratio is indeed closed to 1:1, their total mass closed to 0.2 kg. (Sender, Fuchs and Milo, 2016), and the number of genes encoded by microbes (~9,879,896 genes) is now up to 450x time the size of our genome (~22,000 genes) (Collins *et al.*, 2004; Li *et al.*, 2014). The human genomes have about 99.9% similarity (Wheeler *et al.*, 2008), however it has been shown that all our microbial genes (microbiome) can reach 80-90% differences (Turnbaugh *et al.*, 2009). Previous studies of the human gut (Qin *et al.*, 2010) from the European MetaHIT consortium, found between 1,000 and 1,150 non-redundant and prevalent bacterial species, harboring each individual at least 160 taxa. However, other studies less conservatives estimate the number of bacterial species to be up to 15,000-40,000 individual members (Frank and Pace, 2008).

The human body is first colonized at birth during the vaginal delivery and has a potential impact on human health and disease(Gensollen *et al.*, 2016). When the delivery is vaginal, the infant acquires a microbiota more resembling to that of the mother's vagina dominated by genera *Lactobacillus* and *Prevotella*, whereas if the delivery is c-section the microbiota resembles more that present in the skin, dominated by genera *Propionibacterium*, *Staphylococus* and *Croynebacterium* (Dominguez-Bello *et al.*, 2010; Jakobsson *et al.*, 2014). In the gut, the microbiota is dominated by bacterial phyla such as "Firmicutes" and "Bacteroidetes" which are found throughout the intestinal tract with other microbes such as archaebacteria, viruses (Breitbart *et al.*, 2008) and eukaryotes like fungi.

The gastrointestinal tract (GI) harbors most of our bacteria, with the colon being the area with highest density. The alterations of the GI microbiota can affect many parts of the human body as illustrated in the Figure 1.1.

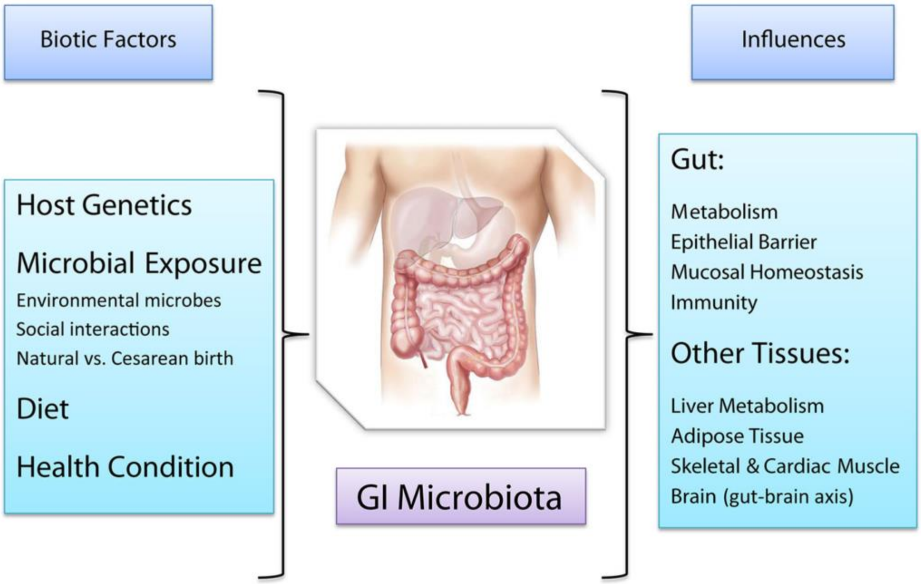


Figure 1.1 Alterations of the GI microbiota.
Factors affecting the human GI microbiota and host functions affected, either directly or indirectly, by the GI microbiota (Selber-Hnativ et al., 2017)

The bacterial diversity in the human digestive tract was extensively investigated in a previous study (Stearns *et al.*, 2011), and they found that the highest bacterial richness and phylogenetic diversity was located in the mouth. The phylogenetic variability between subjects was higher than there was between sample sites from within each gastrointestinal location (e.g. mouth, large intestine). These results were coherent with other studies (Caporaso *et al.*, 2011). It has been shown that the GI tract is a changing ecosystem where adaptable microbial communities are replaced continually by functionally similar communities. These observations strengthen the idea that the highest diversity of bacteria found in the mouth, which represents the entry point of the GI tract, is “filtered” or selected in lower habitats of the GI tract, decreasing the diversity as these organisms pass through.

The role of the microbiota is crucial in the human homeostasis, and its composition and activity directly affect the metabolism of our organism. Its main commitment is to produce a broadly spectrum of metabolites that are not directly obtained by the GI tract itself, hence contributing to the human nutrient metabolism. They pose a constant threat of invasion owing to their total numbers and the large intestinal surface area. The intestinal immune system maintains constant homeostatic interactions with the current resident microbiota (Garrett, Gordon and Glimcher, 2010; Hooper and MacPherson, 2010)

The microbial community composition varies throughout the life, and is shaped by the genetic background of the host, diet and the health status (Ottman *et al.*, 2012).

The stools or faeces conform the targets of most of the studies of the GI tract due to the easy access and the fact that they contain a huge number of microbes, facilitating the recovery of microbial nucleic acids molecules for metaomics analyses. The diversity of microbes is lower in high taxonomical ranks (dominated by Firmicutes and Bacteroides at phylum level), and higher at lower taxa classifications, like species or strains (Grice and Segre, 2012).

Coupled with the term microbiota, the term microbiome refers to the genes and genomes of the microbes forming the microbiota. The microbiome is also referred to as our second genome in some studies (Grice and Segre, 2012), and represents the “material” used for the study of the microbiota using molecular approaches . Each microbiome differs in composition and functions in each body site (Peterson *et al.*, 2009) and evolves over the time (Ottman *et al.*, 2012), however, some studies have been able to associate geographical regions and lifestyles with the microbiota of their healthy population (Andersson *et al.*, 2008; De Filippo *et al.*, 2010; Claesson *et al.*, 2012; Qin *et al.*, 2012; Yatsunenko *et al.*, 2012; Tyakht *et al.*, 2013). These studies suggest that regardless of its variability, there is stability in terms of composition over periods of life and regional lifestyles. As other researchers point out, there is increasing evidence that individuals actually share a “core microbiome” rather than “core microbiota” (Ursell *et al.*, 2012).

The microbial functions play a key role in the digestion of the nutrients we obtain from food. Without them many nutrients cannot be broken or discomposed by our intestinal tract itself, and thus we could not get benefit of them. Furthermore, the gut microbial ecosystem has been implicated in many diseases: related to brain-gut axis dysfunction (Cryan and Dinan, 2012), obesity (Turnbaugh *et al.*, 2006), IBD (Frank *et al.*, 2007; Sokol *et al.*, 2008; Kaser, Zeissig and Blumberg, 2010), liver (Chassaing, Etienne-Mesmin and Gewirtz, 2014), diabetes (Qin *et al.*, 2012), and atherosclerosis (Koeth *et al.*, 2013) among others.

It is worth to mentioning that the study of the human microbiota is not new, it started early in 1680s with Antonie van Leewenhoek, when he found differences in microbes between two distinct body habitats. The current novelty consists on the ability to use the new molecular techniques coupled with bioinformatic and biostatistical analysis that bring new insights into the mechanisms in which the microbial communities are involved in maintaining a healthy status or in the onset and/or perpetuation of diseases. (Ursell *et al.*, 2012).

In 2011 the Human Microbiome Project (HMP) analyzed one of the largest cohorts of healthy individuals that established an initial characterization of what we consider a healthy microbiome in western populations. Distinct body sites (i.e. gut, skin, vagina) where analyzed, and they observed that the microbiota differed notably in terms of diversity. Within-subject variation was lower than between-subjects, and the microbial community of everyone was stable over time. However, they could not observe a common group of taxa among all body habitats. In terms of functions, the metabolic pathways were stable among individuals, whereas in terms of taxonomical diversity it was variable (Huttenhower *et al.*, 2012). The source of this high diversity is still not clear, but it is thought that factors like diet, environment and genetics are implicated.

Over the last decade, the human microbiome has been the focus of important international consortia such as the Human Microbiome Project (HMP, 2008-2017) and the Integrative Human Microbiome Project (iHMP, 2018 onwards), both a NIH (United States National Institutes of Health) initiative, and Metagenomics of the Human Intestinal Tract (MetaHIT, 2008-2012), an European consortium. These consortia have deposited catalogues of microbial genes in an unprecedented amount (Huttenhower *et al.*, 2012; Li *et al.*, 2014). If we pay attention to the number of papers mentioning the word “microbiota” or “microbiome” in the medical literature within the NCBI-PubMed database, we can see as the publications have been increasing exponentially over the last fifteen years.

1.2 Inflammatory Bowel Disease (IBD)

The inflammatory bowel disease (IBD), refers to a group of chronic intestinal diseases that produce inflammation of the gut, and includes two different types: Crohn’s Disease (CD) and Ulcerative Colitis (UC). CD involves intestinal inflammation that could affect different sites of the entire GI tract, but concerns more frequently the terminal ileum and colon (small and large intestine), although it can also affect the other parts of the GI tract (i.e. mouth, esophagus, and stomach), whereas UC is limited to the mucosa and submucosa of the colon (epithelial lining of the gut) (Huang *et al.*, 2014)(see Figure 1.2).

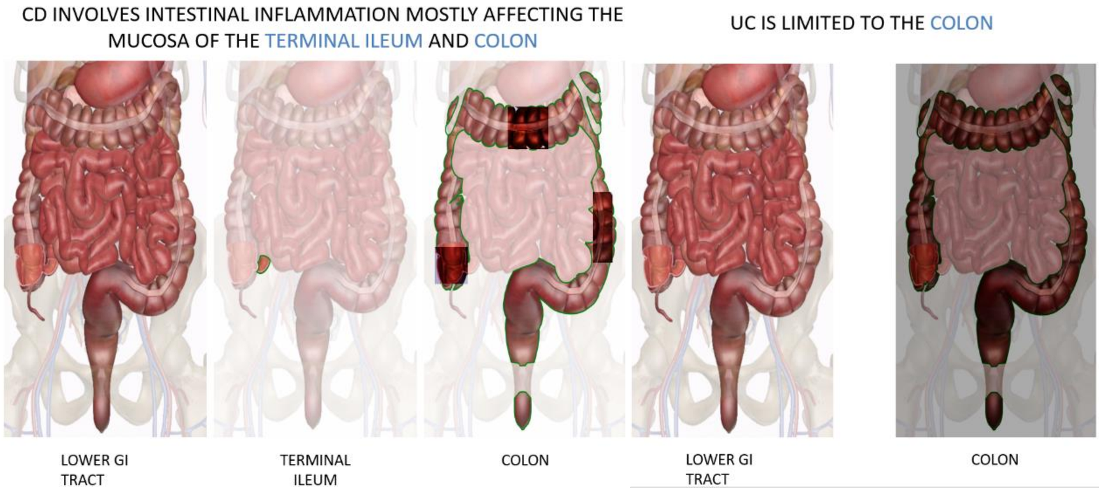

CD INVOLVES INTESTINAL INFLAMMATION MOSTLY AFFECTING THE MUCOSA OF THE TERMINAL ILEUM AND COLON
UC IS LIMITED TO THE COLON
LOWER GI TRACT
TERMINAL ILEUM
COLON
LOWER GI TRACT
COLON

Figure 1.2 Parts of the bowel affected in patients with CD or UC

The common symptoms of patients with CD and UC are: abdominal pain and cramping, bleeding ulcers and recurring diarrhea (Fakhoury *et al.*, 2014). Both types of diseases show a variable course of activity, followed by very few sporadic or induced remissions of the intestinal damages and spontaneous relapsing attacks (Manichanh *et al.*, 2012)

Inflammatory bowel disease was a very rare disease in the beginning of the last century, but during half of the last century its incidence is increasing extraordinarily and in 2015 IBD was afflicted an estimate of 3.6 million people in

Europe and USA (Meyer *et al.*, 2012; Lee and Maizels, 2014; Kaplan, 2015; Ng *et al.*, 2017) and keeps increasing and expanding to other countries (see Figure 1.3), almost reaching 1% of the population and leading to the assumption that will become a worldwide epidemic (Manichanh *et al.*, 2012).

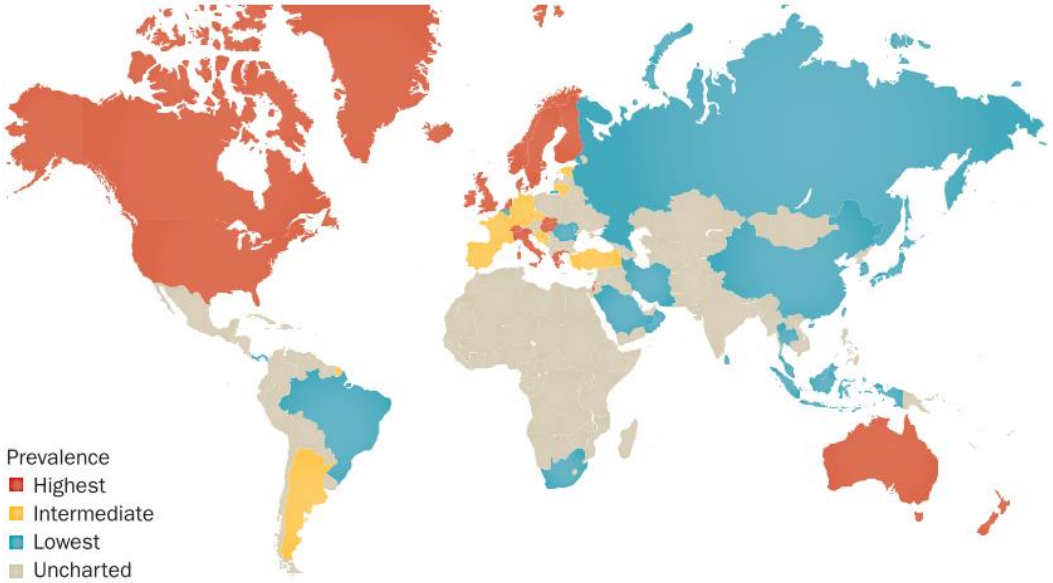


Figure 1.3 Prevalence of IBD in 2015 (Kaplan, 2015)

Although the aetiology is still uncertain, it has been linked to environmental factors (i.e. diet, antibiotic use, social status and microbial exposure among others) that may trigger immunological responses that inflame and damage the GI (Lee and Maizels, 2014). It is characterized by a dysbiosis or imbalance of the microbiota, and increasing evidence suggests that it may be linked to the genetic of the host (Meyer *et al.*, 2012). Nonetheless, in other studies they calculated that the genetic predisposition of the host in relation to IBD only contributes the 23% in CD and 16% in UC (Peloquin *et al.*, 2016). This opens up a window to explore other factors that might play an important role in the development of the disease.

The microbiota has been linked with the regulation of the mucosal immune system and has been recognized as the main player that leads to chronic intestinal inflammation. The complexity of the gut microbiota makes difficult the understanding of the relationships between microbes and their relationship with the host.

The microorganisms have been co-living with us since the beginning of our species, their importance cannot be neglected since our physiology depends on their symbiotic relationship. In contrast, some bacteria have been implicated in the pathogenesis of many inflammatory diseases like IBD. This places the microbiota as a key factor for maintaining the homeostasis of the mammalian immune system (Hill and Artis, 2010).

The reduction in diversity of bacterial communities in IBD patients compared to healthy individuals have been previously described (Manichanh *et al.*, 2006). And it is known that this dysbiosis is accompanied by dramatic productions of cytokines (proteins involved in cell signaling), T cell (subtype of white blood cell playing a central role in immune response) activations, and IgC (immunoglobulin) antibody response to intestinal bacteria (Duchmann *et al.*, 1995; Macpherson *et al.*, 1996). In another study (Sarrabayrouse *et al.*, 2015), they identified, in humans, a mechanism by which the gut microbiota can affect the gut homeostasis via the induction of DP8α Tregs (type of T cell).

1.3 Approaches to the study of the microbiome

A couple of decades ago the advent of new and fast computational advances has revolutionized the way we communicate and process data. Computers are now present everywhere and are paramount to perform complex tasks. The biology field was also impregnated of this revolution, and since the first methods to sequence the DNA by chain termination techniques (Sanger sequencing or first generation), other techniques like sequencing by DNA synthesis (Next Generation Sequencing, High Throughput Sequencing or second generation) have had a great impact in the research area. NGS technologies have allowed sequencing DNA molecules at a very low-cost and have thus boosted the use of meta-omic[1] approaches to study microbial communities. Now, the major challenge is to create, develop and standardize bioinformatic tools able to process this torrent of data produced by massive parallel reactions, with the use of multiprocessing, multithreading, and computer clusters.

The old-school methods to study the microbial populations consisted in a bottom-up experimentation, a hypothesized “culprit” microorganisms or genes are selected as candidates and studied individually, either by culture-based methods or by direct observations using imaging technology. This approach uses supervised methods of analysis. Now, the “omics” era via high-throughput sequencing opens the possibility to perform research top-down by using massive molecular content of microbial communities with no prior hypothesis on what/who is behind the curtains. This approach uses unsupervised methods. The molecular content primarily can be either DNA (metagenomics), RNA (metatranscriptomics), proteins (metabolomics) or metabolites (metabolomics) (see Figure 1.4) (Huang *et al.*, 2014).

[1] The “meta-“ suffix refers to the concept of “going further” (i.e. not limited to the study of one organism, but to the study of the relationships/interactions of a group of organisms) and the “-omic” suffix indicates the possibility to study a large number of biological material (i.e. genes, proteins, transcripts, …); metagenomics for instance means the study of the genes of several organisms at once and the relationships they might have. Meta-omic approaches help to understand complex microbial communities as a whole.

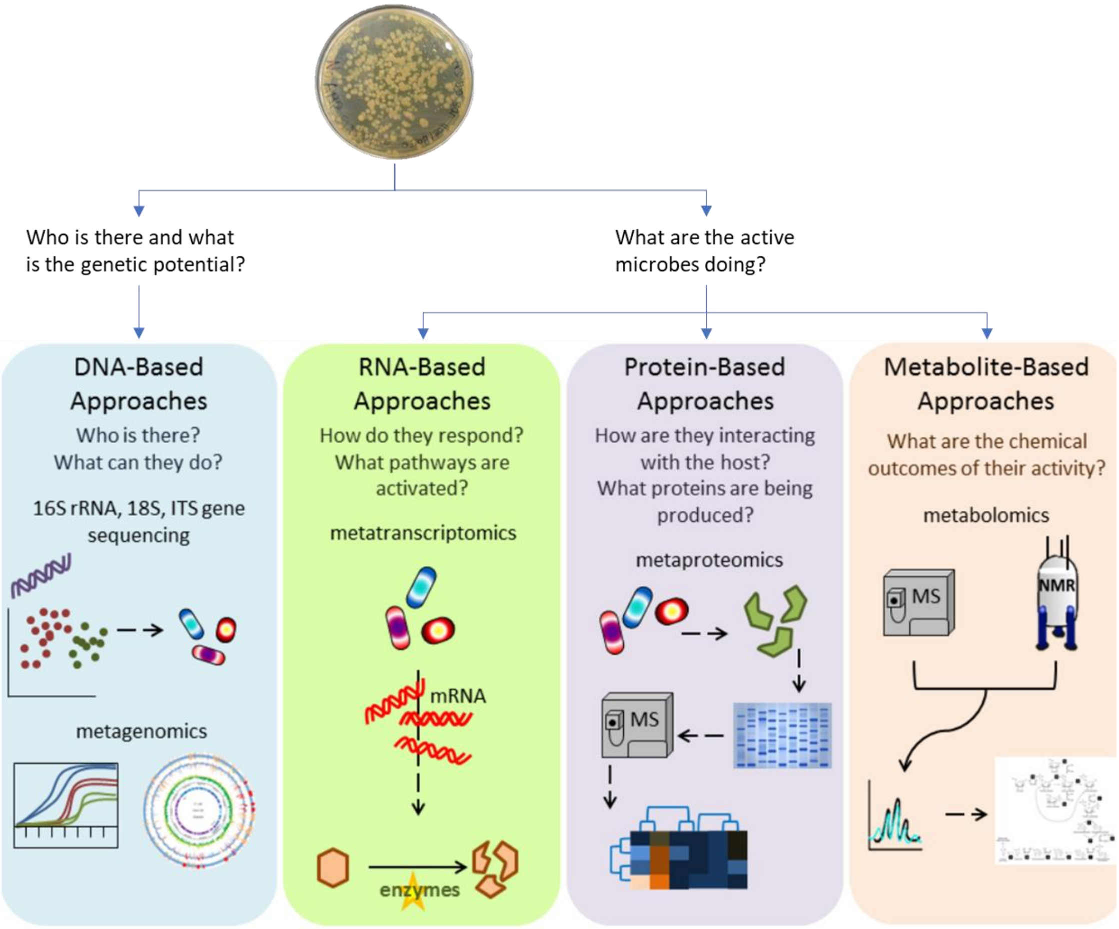

Who is there and what is the genetic potential?
What are the active microbes doing?
DNA-Based Approaches
Who is there? What can they do?
16S rRNA, 18S, ITS gene sequencing
metagenomics
RNA-Based Approaches
How do they respond? What pathways are activated?
metatranscriptomics
mRNA
enzymes
Protein-Based Approaches
How are they interacting with the host? What proteins are being produced?
metaproteomics
MS
Metabolite-Based Approaches
What are the chemical outcomes of their activity?
metabolomics
MS
NMR

Figure 1.4 Molecular approaches to study the microbiome.
(National Academies of Sciences, Engineering, and Medicine, 2018). The culture-plate picture (https://vdsstream.wikispaces.com/ChristinaP) of unknown author is under license CC BY-SA (https://creativecommons.org/licenses/by-sa/3.0/)

1.4 NGS Technology

The first sequencing technologies based on DNA chain termination were developed by Frederick Sanger and his colleagues at the end of the seventies. That advance changed biology by offering new tools to understand the genes and genomes from a molecular point of view. The common term used to reference this first generation of sequencing machines was coined as “Sanger sequencing”, being Applied Biosystems (ABI) the first company to produce commercial sequencers implementing that methodology. The main drawback of these sequencers were the expensive cost and the number of sequenced bases per run, up to 96,000 bp in the Sanger ABI 3730xl model (Rhoads and Au, 2015); as a reference, the human genome is around 3.3 billion bp long (Kchouk, Gibrat and Elloumi, 2017). However, by that time there was no other cheaper alternative, and the first generation survived for 30 years. The first genome that could be sequenced was the phiX174 enterobacteria phage with a size of 5,374 base pairs (bp) (Kchouk, Gibrat and Elloumi, 2017).

Table 1 Sequencing platforms features.

Method	Generation	Read length (bp)	Single pass error rate (%)	No. of reads per run	Time per run	Cost per million bases (USD)
Sanger ABI 3730×l	1st	600-1000	0.001	96	0.5-3 h	500
Ion Torrent	2nd	200	1	8.2 × 107	2-4 h	0.1
454 (Roche) GS FLX+	2nd	700	1	1 × 106	23 h	8.57
Illumina HiSeq 2500 (High Output)	2nd	2 × 125	0.1	8 × 109 (paired)	7-60 h	0.03
Illumina HiSeq 2500 (Rapid Run)	2nd	2 × 250	0.1	1.2 × 109 (paired)	1-6 days	0.04
SOLiD 5500×l	2nd	2 × 60	5	8 × 108	6 days	0.11
PacBio RS II: P6-C4	3rd	1.0-1.5 × 104 on average	13	3.5-7.5 × 104	0.5-4 h	0.40-0.80
Oxford Nanopore MinION	3rd	2-5 × 103 on average	38	1.1-4.7 × 104	50 h	6.44-17.90

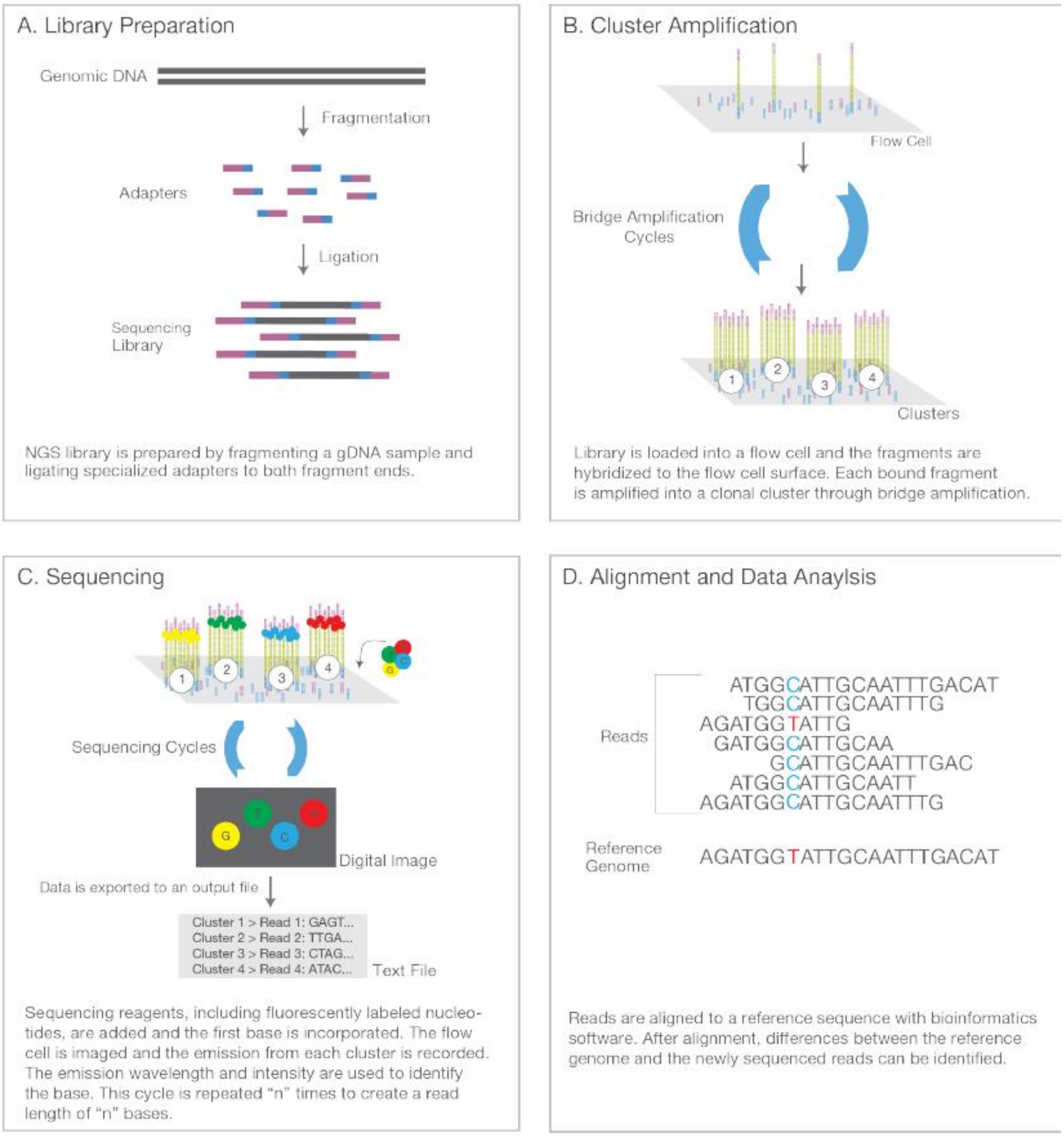
Performance of different sequencing platforms between generations (Rhoads and Au, 2015)

By means of Sanger sequencing in 2004, and after 15 years of work, the human genome was completely sequenced in the Human Genome Project (HGP) (Collins *et al.*, 2004; Jaszczyszyn *et al.*, 2014). The efforts in terms of time, costs and resources exhibited the evidence that faster, high-throughput and cheaper sequencers were required in future projects to overcome the huge limitations of these technologies in smaller research groups. General efforts from institutions succeeded to reduce the costs, and after ten years, by 2015 the cost was close to 1,000$, a reduction of 100-fold compared to the 100 million dollars needed for the HGP (Schloss, 2008; Jaszczyszyn *et al.*, 2014; Kchouk, Gibrat and Elloumi, 2017; KA., Wetterstrand, 2018).

That framework empowered companies to develop better sequencers, and the second generation appeared in 2005 with the 454 Life Sciences Genome Sequencer FLX based on pyrosequencing (a methodology based on “sequencing by synthesis”). This was the first of a new variety of sequencers also referred to as Next Generation Sequencing (NGS) technologies. They were mainly characterized by the increase in the number of sequenced DNA bases per run, from thousands of bases in the first generation to millions of parallel reactions that increased the sequencing throughput enormously. Another important feature was the use of in-vitro DNA amplification (library preparation and PCR) instead of the common in-vivo bacterial cloning amplification methods used in Sanger. If the sequencing of the human genome took over 15 years in the HGP project, using the 454 GS FLX sequencer they required only two months (Kchouk, Gibrat and Elloumi, 2017). The term High-Throughput NGS is also used to denote this technology.

The typical NGS workflow consists on the initial extraction of the genomic DNA from a single organism or an entire population. This DNA is then chopped into small fragments between 50 to 500 nucleotides suitable for the sequencing process. The fragments are later required to be attached to adapters in both end of the fragments in order to be manipulated by the sequencers. The main purpose of the adapter is to stick the fragments in a solid surface and to allow the sequencing primers to bind to the sequence to read it. The next step is a PCR (Polymerase Chain Reaction) amplification of the sequences, this is required to make the “signal” stronger enough to be detectable by the sequencer. Last, the sequencer reads each nucleotide and produces the so called “reads” from each sample which represent just a conversion of nucleotides into computer files in the form of A, C, G or T letters (N in case of ambiguity)(Figure 1.5).

The sequencing of the library can be done in one-sense (single-end reads) or in both senses of the fragment (pair-end reads). In the later, depending on the library preparation kit used for the sequencer only one of the two reads will dictate the stranedness. Usually “read1” dictates strand (i.e. ScriptSeq™ from Epicentre; employs a directional protocol tag-based), but in some other kits the “read2” might be the one dictating (i.e. Illumina® TruSeq® Stranded Total RNA; employs the dUTP protocol). See Figure 1.5 and Figure 1.6.



NGS library is prepared by fragmenting a gDNA sample and ligating specialized adapters to both fragment ends.

Library is loaded into a flow cell and the fragments are hybridized to the flow cell surface. Each bound fragment is amplified into a clonal cluster through bridge amplification.

```
Cluster 1 > Read 1: GAGT...
Cluster 2 > Read 2: TTGA...
Cluster 3 > Read 3: CTAG...
Cluster 4 > Read 4: ATAC...
```

Sequencing reagents, including fluorescently labeled nucleotides, are added and the first base is incorporated. The flow cell is imaged and the emission from each cluster is recorded. The emission wavelength and intensity are used to identify the base. This cycle is repeated “n” times to create a read length of “n” bases.

```
Reads
      ATGGCATTGCAATTTGACAT
        TGGCATTGCAATTTG
  AGATGGTATTG
    GATGGCATTGCAA
          GCATTGCAATTTGAC
      ATGGCATTGCAATT
  AGATGGCATTGCAATTTG

Reference Genome
  AGATGGTATTGCAATTTGACAT
```

Reads are aligned to a reference sequence with bioinformatics software. After alignment, differences between the reference genome and the newly sequenced reads can be identified.

Figure 1.5 Overview of the NGS workflow of basic chemistry and bioinformatic steps.
(A) Library preparation, (B) cluster generation, (C) sequencing, (D) data analysis and alignment of reads to reference genome if reads belong to a known organism (Illumina, 2012)

The negative counterpart of the NGS technology remains on its limitation to sequence small parts of the DNA, i.e. a single molecule cannot be sequenced. Thus, the production of millions of short reads needs to be later assembled into a genome by using specific algorithms and high computer resources. The only way to obtain a higher genome accuracy is to get a higher coverage of DNA sections, i.e. produce more reads that may overlap a section and support the reliability of the read sequence. All this inherently comes with the requirement of high-storage capacity and high-memory consumption. Another limitation that comes with NGS is the Polymerase Chain Reaction (PCR) bias introduced during the amplification of the library (Jaszczyszyn *et al.*, 2014; Kchouk, Gibrat and Elloumi, 2017).

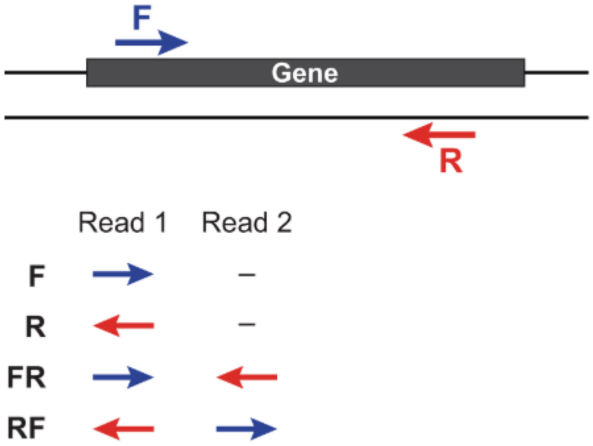


Figure 1.6 Strand-specific libraries.
The left and right sequencing reads are depicted according to orientations relative to the sense strand of a transcript sequence. Four configurations can be used depending on the library type (F, R, FR, RF). The "RF" (reverse is the "Read1" and forward the "read2") corresponds to the "Illumina® TruSeq® Stranded Total RNA" library (dUTP second strand protocol) used for RNA-Seq *(Haas* et al.*, 2013)*.

The third generation of sequencing technologies came in 2010 with Pacific Biosciences (PacBio RS) promising longer reads ranging from 1,300bp to 13,500bp at low sequencing costs, eliminating the need for the PCR amplification, and at a faster run-time. This increase in read lengths solved the problems of NGS with short reads, that made difficult to identify repetitive areas of complex genomes, making this generation appropriate to recognize de novo genomes without the need of reference genome to perform the assembly. Regardless the higher error rates, if NGS had a maximum of 1% of sequencing error, this generations comes with around a 15% error rate, they were able to produce consensus assemblies with error rates comparable to the first and second generation (Jaszczyszyn *et al.*, 2014; Kchouk, Gibrat and Elloumi, 2017; Mardis, 2017).

1.5 Metagenomics and Metatranscriptomics

The microbiota has a direct impact on the health of the environmental niche where it resides, a dysbiosis, or microbial imbalance, produced by a modification in the diet of the host for instance, could affect drastically the products consumed and produced by the community which in turn can be dangerous for the health of the host (Foxman and Martin, 2015; Aguiar-pulido *et al.*, 2016). The study of the DNA of the microbes residing in the microbiota, cultivated or uncultivated, can be studied using metagenomics. This approach relies on the microbial DNA sequencers to obtain a picture of the taxonomical and potential functional profile of a sample. However, while metagenomics aims to know who is there by identifying partially/completely the genome of the microorganisms, or by targeting the 16S ribosomal RNA marker gene and using it roughly as a car plate, metatranscriptomics focus on the RNA material of the community. This approach is able to capture those microorganisms that are “*active*”, i.e. are transcribing genes from DNA to RNA (transcription), by identifying the 16S rRNA gene at RNA level instead of DNA (to differentiate the 16S rRNA gene at RNA level from the DNA form, the latter is usually referred to as 16S rDNA). This technique allows to identify not only the *“active”* microorganisms but also the *“active”* functions that are being activated in the niche by the different microorganisms, i.e. those genes transcribed by means of the RNA polymerase that will be later translated to functional gene products (gene expression). If the product is a protein, the process is called protein synthesis.

While metatranscriptomics can directly inform upon active functional profile, it cannot answer well to the question of which functions or genes are differentially expressed between different states. In those situations, the underlying microbial community taxa must be taken in consideration. Thus, when a differentially expressed function is met between a two conditions setup, it is difficult to interpret biologically that difference, is it due to a change in the expression level of the two conditions from the same group of microorganisms (taxa)? Or, since microbes compete or cooperate for nutrients, feed, grow and reproduce, is it due to a change in taxa abundance? (Franzosa *et al.*, 2014; Morgan and Huttenhower, 2014; Aguiar-pulido *et al.*, 2016; Bashiardes, Zilberman-Schapira and Elinav, 2016; Klingenberg and Meinicke, 2017; Abu-Ali *et al.*, 2018; Mehta *et al.*, 2018). Those questions are very difficult to address using only RNA data, it is necessary to normalize by the DNA copy number in order to have an unbiased interpretation of results, here is where metagenomics becomes necessary, and combined with metatranscriptomics will provide a better snapshot of the activated functions under certain conditions (Shi *et al.*, 2011; Giannoukos *et al.*, 2012; Morgan and Huttenhower, 2014). As it has been shown in a previous study, it is important to remark that the metatranscriptomes vary more within individuals than metagenomes (Franzosa *et al.*, 2014; Knight *et al.*, 2018).

The following chart (Figure 1.7) clearly illustrates that, despite new research is being held every new year, metatranscriptomic analyses still remains poor compared to transcriptome analyses of a species or single-cell.

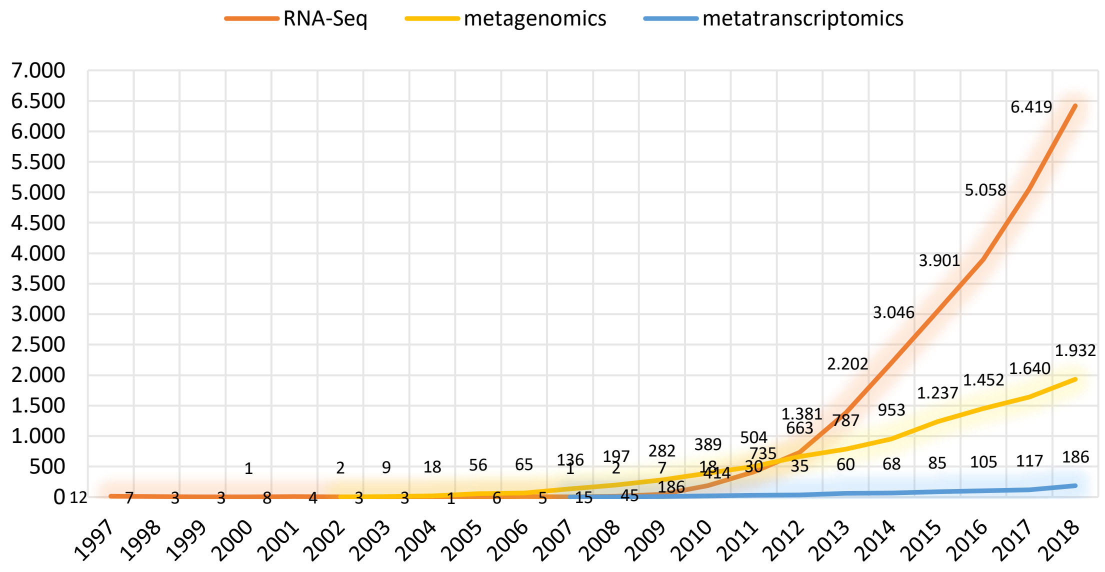


Figure 1.7 Comparison of publications per year in RNA-Seq, metagenomics and metatranscriptomics.

Chapter 2.

Hypothesis and objectives

2 Hypothesis and objectives

Since complex microbial communities, such as the human gut microbiota, are mostly reluctant to culture methods, omics approaches using high-throughput sequencing techniques have been very useful as alternatives to characterize their composition and functions. However, to date, production of massive data challenges the modern computers to effectively process this data and obtain a reliable characterization of the microbial communities.

Hypothesis

The hypothesis of this PhD was based on the assumption that characterizing the active functions of the microbiome will help understanding not only the role of the gut microbiota in a healthy state but will also unravel its implication in the development and perpetuation of intestinal disorders.

Main objective

Taking advantage of high-throughput sequencing technologies and an in-house developed bioinformatic pipeline, will allow to comprehensively determine which microbial members and which functions of the gut microbial community are active and associated with health, FBD (Functional Bowel Disorders) and IBD (Inflammatory Bowel Disorders).

Secondary objectives

The aim of this dissertation is:

-

1. First, the development of a reliable and efficient pipeline to perform metatranscriptomic analysis using the power of multi-threading computers. Its modular design must allow an easy interchangeability or improvement of any of the stages involved in the analyses.

2. Then, this pipeline will be applied to characterize the human microbiome in health and disease states.

The work produced to achieve the first of the secondary objectives lead to the publication of one article in “Nature Scientific Reports” (2016):

> Martinez X, Pozuelo M, Pascal V, et al. MetaTrans: an open-source pipeline for metatranscriptomics. *Sci Rep*. 2016;6:26447. doi:10.1038/srep26447.

Chapter 3.

Materials and methods

3 Materials and methods

3.1 Pilot study

3.1.1 Ethics statement

The methods were carried out in accordance with the approved guidelines. All the experimental protocols were approved by the Institutional Review Board of the Vall d'Hebron Hospital (Barcelona, Spain). Subjects provided their written informed consent to participate in this study.

3.1.2 Design and samples collection protocol

The pipeline was designed to perform two types of RNA-Seq analyses, namely those addressing 16S rRNA taxonomy and gene expression. To test the present metatranscriptomic pipeline, we analyzed synthetic mock communities and twelve fecal samples collected from eight individuals obtained from a previous study (Manichanh *et al.*, 2013) and from an unpublished one. For four individuals, before and after a flatulogenic diet challenge of three days, stool samples were collected, and intestinal volume of gas was measured.

To test the pipeline with RNA-Seq newly generated, RNA sequencing was performed in two types of experimental designs, named:

- “total RNA” (eight samples) and
- “rRNA removal” (four samples)

experiments from here onwards.

The objective of the “total RNA” sequencing experiment was to recover both the functional and taxonomic profile of each active microbial community in an unbiased manner. This experiment was performed on eight stool samples from four individuals in two time points that were collected in a previous work (Manichanh *et al.*, 2013). As shown in previous studies (McNulty *et al.*, 2011; Huttenhower *et al.*, 2012; Cotillard *et al.*, 2013; Manichanh *et al.*, 2013; Tap *et* *al.*, 2015), the diet can have a great impact in the functional response of the microbial community. Thus, to detect functional variations for each participant, samples were collected before and immediately after three days of a flatulogenic diet. In the present study, we believe that combining 16S rDNA, 16S rRNA and mRNA data can provide a new perspective of the factors involved in the origin of flatulence.

Stools were collected from four participants - two healthy and two diagnosed with FBD (Functional Bowel Disorders) and complaining of excessive gas evacuation (flatulence). The subjects were instructed to follow their usual diet for 3 days and to consume a diet rich in fermentable residues for another 3 days, during which each meal (breakfast, lunch, dinner) included at least one portion of the following: (a) bread, cereals or pastries made of whole wheat or corn; (b) beans, soya bean, corn, broad beans, or peas; (c) brussels sprouts, cauliflower, broccoli, cabbage, celery, onion, leek, garlic or artichoke; and (d) banana, fig, peach, grapes or prunes. The volume of intestinal gas was measured as previously described (Serra *et al.*, 2002; Hernando-Harder *et al.*, 2010). The gas collection tests were conducted before and after the flatulogenic diet using a rectal balloon catheter (20 F Foley catheter, Bard, Barcelona, Spain) connected via a line without leaks to a barostat, and the volume was continuously recorded (Table 2).

Table 2 Volume of intestinal gas.

Sample name	Before diet (ml)	After diet (ml)
#1	284	446
#2	410	1621
#3	167	967
#4	135	573

Recorded using a rectal balloon catheter connected via a line without leaks to a barostat.

Two methods of taxonomic analysis were compared, one using 16S rRNA extracted from the “total RNA” experiment and the other using 16S rDNA V4 amplicons of the same samples obtained from a previous study (Manichanh *et al.*, 2013).

On the other hand, the objective of the “rRNA removal” experiment was to test how the rRNA depletion step would increase the recovery of number of expressed genes. This experiment was performed on four additional stool samples obtained from four individuals.

Stools were collected from four participants - two healthy subjects and two patients with CD (Crohn’s Disease) a type of IBD (Inflammatory Bowel Disease). The stool collection protocol involved providing participants with an ice bag containing an emesis basin (Ref. 104AA200, PRIM S.A, Spain), a 50-mL sterile sampling bottle (Deltalab, Spain), a sterile spatula (Deltalab, Spain), and gloves during their visit to the laboratory. For the purpose of stool collection, participants were instructed to do the following at home: 1) use the emesis basin provided to collect the stool; 2) after the deposit, transfer it to the sampling bottle, ensuring proper homogenization; and 3) take the sampling bottle to the laboratory within the first three hours after deposit or, if not possible, store it in the home freezer (-20 °C) and take it to the laboratory properly surrounded by frozen gel blocks as soon as possible. Once in the laboratory, the samples were stored at -80 °C until processed.

3.1.3 Genomic RNA extraction

Using the twelve collected samples, the total RNA was extracted, performed an rRNA removal step in a set of four of them in the “rRNA removal” experiment, and prepared cDNA libraries for paired-end sequencing to increase the read fragment and improve the read mapping (Li, 2013) by Illumina machines.

For the extraction of the total RNA was used the protocol described in a previous study (Cardona *et al.*, 2012). Briefly, 250 mg of fecal sample was mixed with 500 µl of TE buffer, 0.8 g of Zirconia/silica Beads, 50 µl of SDS 10% solution, 50 µl of sodium acetate, and 500 µl of acid-phenol. Physical disruption was achieved using a FastPrep apparatus (FP120, 101Thermo). Following centrifugation of the lysate, nucleic acids were recovered from the aqueous phase and re-extracted with chloroform:isoamylalcohol. DNA was selectively digested, and RNA was purified using the RNeasy® mini kit (Cat. No. 74104, Qiagen), following the manufacturer's instructions. Total RNA for an equivalent of 1 mg of each fecal sample was quantified using a Nanodrop ND-1000 Spectrophotometer (Nucliber) while the quality was assessed using a RNA6000 Nano chip (total RNA) in an Agilent 2100 Bioanalyzer. RNA quality was determined by the RNA integrity number (RIN), which is calculated from the relative height and area of the 16S and 23S RNA peaks and follows a numbering system from 1 to 10, 1 being the most degraded profile and 10 the most intact. The average RIN number obtained was 6.8, with values ranging from 6.3 to 7.4.

3.1.4 rRNA removal and cDNA synthesis and sequencing

For the "total RNA" experiment, total RNA of eight samples was subjected to fragmentation of 50 ng of RNA molecules; after, complementary DNA (cDNA) of the RNA was synthesized using the ScriptSeq™ v2 RNA-Seq Library Preparation Kit (directional RNA-Seq) from Epicentre (an Illumina® company) with random hexamers, and incorporation of Illumina platform-specific 3′ sequencing tag (tag based, where Read1 dictates strandedness). The multiplexing index was added through 12 cycles of PCR performed using the FailSafe™ PCR Enzyme Mix (Epicentre Biotechnologies, #FSE51100) followed by AMPure XP Purification (Agencourt, Beckman Coulter). Each library was sequenced as paired-end 76-bp reads on the Illumina HiSeq 2000 platform (Centre Nacional d'Anàlisi Genòmica, CNAG, Barcelona, Spain) and produced 16 files (8 paired-end), which generated a total of 24 Gbp.

For the “rRNA removal” experiment, total RNA of four samples was subjected to an rRNA depletion procedure using the Ribo-zero Magnetic kit according to the manufacturer’s instruction (Epicentre, an Illumina® company). The samples were then subjected to fragmentation of the remaining RNA molecules; after, complementary DNA (cDNA) of the RNA was synthesized using the TruSeq® Stranded mRNA Library Preparation Kit from Illumina® (dUTP based, where Read2 dictates strandedness) where the poly-A selection method for ribosomal reduction was discarded. Each library was sequenced as paired-end 101-bp reads on the Illumina HiSeq 2000 platform (Centre Nacional d'Anàlisi Genòmica, CNAG, Barcelona, Spain) and produced 8 files (4 paired-end), which generated a total of 22 Gbp.

All sequenced samples are available at NCBI SRA project id: PRJNA295252.

3.1.5 Bioinformatic analysis

For the computational analysis, a computer server was used with 2 cores of 8 CPUs each (allowing up to 32 threads enabling Intel® Hyper-Threading technology), 128GiB of RAM, 8TiB of free space. The server used a linux 64bits operative system Ubuntu 14.04 64-bits LTS (Trusty) with the kernel 3.13.0-100-generic x86_64.

3.1.5.1 Sequence analysis steps

From the Illumina platform, we obtained paired-end reads in FASTQ format (CASAVA 1.8, Phred + 33) separated into distinct files for each single-end read and for each sample. The analysis was performed in four major steps described as such in Figure 3.1: filtering, sorting, and functional and taxonomic annotations. The backbone of the pipeline was written in POSIX shell and the internal scripts were written in POSIX, Python, R and AWK languages.

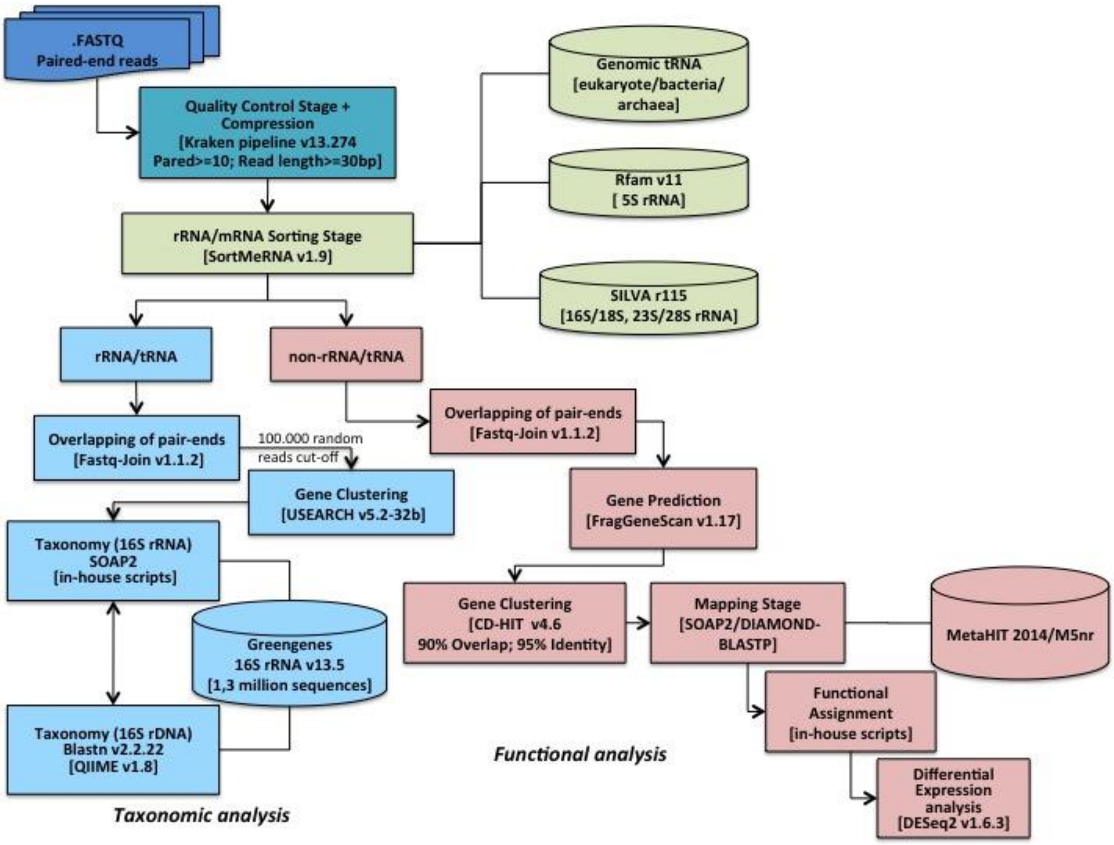


Figure 3.1 Flow diagram of the metatranscriptomic pipeline.
The raw paired-end reads were subjected to quality control and adjustment using the FastQC tool and Kraken pipeline (turquoise boxes). The rRNA/tRNA reads were then separated from the non-rRNA/tRNA reads using SortMeRNA software (green boxes), for taxonomic (clear blue boxes) and functional analyses (pink boxes), respectively. For the taxonomic analysis, the reads were mapped against the 16S rRNA Greengenes v13.5 database using SOAP2. For the functional analysis, the reads were subjected to the FragGeneScan to predict putative genes before being mapped against a functional database (MetaHIT-2014 or M5nr) also using the SOAP2 tool (see methods for details).

3.1.5.2 Trimming and filtering of quality reads

The raw reads were submitted to the quality control report-tool FastQC (Andrews, 2010), which allows evaluation of the quality of the reads and selection of the most appropriate filtering parameters, such as the per base N content, the read length, and the per sequence quality score, for downstream quality control analysis of the reads. The Kraken pipeline (Davis *et al.*, 2013) was then used to recover quality reads on the basis of the FastQC report. This set of programs is based on efficient multi-threading and a complete set of tools structured in an independent pipeline. They allow not only common cleaning operations such as removal of low-quality reads and filtering of reads with low length, but also Poly-A trimming, N-masked base trimming, collapsing of reads, maintenance of readpairing along the process, and low-complexity filtering, among other features. This pipeline can be adapted to reads obtained from various sequencing platforms. We set up the configuration of the Kraken tools to maintain the link between paired-end reads during the process and to perform two steps. The first, called “reaper”, relied on the call of a fast and flexible tool designed specifically for short-read processing to trim or remove adapters, as well as to test all reads processed against three criteria: trimming cluster-N regions and removing low quality regions (below a Phred score of 10) and reads with a length < 30 nt. The main task of the second step, named “filter”, was to discard reads that had no counterpart and then collapse all identical reads, i.e. duplicated paired-reads. The header of the read was then modified to include the number of copies of each collapsed read. Finally, the collapsed reads obtained were again subjected to FastQC in order to validate their quality. At this point, if the default quality setting does not cover the quality requirements, the parameters can be refined before analyzing more samples.

3.1.5.3 Sorting

After a quality control of reads, to identify those that were clearly non-rRNA/tRNA and therefore potential mRNA, we used an efficient and parallel tool, namely “SortMeRNA” (Kopylova, Noé and Touzet, 2012), which required rRNA databases such as SILVA v115 (Pruesse *et al.*, 2007), Rfam (Burge *et al.*, 2013), and the Genomic tRNA database (Chan and Lowe, 2009). Using these three databases, reads were grouped into various categories, namely 16S/18S-rRNA, 23S/28S-rRNA, 5S-rRNA, and tRNA, respectively. As outputs, SortMeRNA produced a file for each category, and the unclassified reads were saved in a separate file as non-rRNA/tRNA, that is to say, “potential mRNA reads”.

3.1.5.4 Functional annotation

Paired-end reads were generated for each cDNA fragment. As paired-end reads have been shown to recover fewer false positives than single ones (González and Joly, 2013) in differential expression studies, we assembled, when possible, the single end reads before performing gene prediction (Figure 3.1). Thus, reads classified as “potential mRNA reads” by “SortMeRNA” were first subjected to an overlapping step that merged, when possible, the paired-end reads producing a longer read length. This step was performed using the Fastq-Join tool (Aronesty, 2013) with a minimum overlap of 8 bp and a maximum difference of 10%, as proposed in the MG-RAST pipeline (Meyer *et al.*, 2008). The potential mRNA reads file may still contain a number of undesired sequences that do not provide functional information, such as non-coding regions, and should therefore be removed in order to decrease computation time in downstream analysis. For this step, we used FragGeneScan (v.1.17) (Rho, Tang and Ye, 2010) to predict putative genes and discard the rest. This tool was configured with appropriate parameters to work properly with relatively short reads such as those of Illumina. Predicted genes were then subjected to clustering to further reduce the size of the dataset using CD-HIT v4.6 with an identity threshold of > = 95% and a gene overlap of > = 90%. Information on cluster size was then included in the header of all representative reads. Finally, to recover a functional profile for each sample, the potential mRNA reads were mapped against a functional database such as the latest MetaHIT gene catalog (Li *et al.*, 2014) using SOAP2 (Li *et al.*, 2009), with the first match retained. The MetaHIT-2014 database contains functions that were recovered from about 1,250 human gut microbiome samples and that were annotated with the EggNOGv3 (evolutionary genealogy of genes: Non-supervised Orthologous Groups) functional database. In order to use a more general database to analyze for instance other ecosystems than the gut microbiota, we added to the pipeline another database, M5nr and the possibility to use either M5nr or the MetaHIT-2014 database. The (Wilke *et al.*, 2012) database is a non-redundant protein database provided by the MG-RAST server and contains 15.9 million unique proteins and 5.8 million functional annotations from different sources including Integrated Microbial Genomes (IMG), Genbank, InterPro, Kyoto Encyclopedia of Genes and Genomes (KEGG), PathoSystems Ressource Integration Center (PATRIC), Phage Annotation Tools and Methods (Phantome), Reference Sequence (RefSeq), the SEED Project, UniProt Knowledgebase (UniProt). An in-house script was used to take into account cluster sizes and to discard duplicates. Raw abundance matrices were generated and processed using the DESeq2 package (Love, Huber and Anders, 2014) to uncover the most differentially expressed functions. Up- or down-regulated functions were further plotted into metabolic pathways using iPath2 (Yamada *et al.*, 2011).

3.1.5.5 Taxonomic annotation

In the “total RNA” experiment from the paired-end read files previously classified as rRNA/tRNA, the two single reads from the DNA fragment were overlapped using Fastq-Join to increase read lengths and annotation accuracy. From the file containing all overlapped reads for each sample, we randomly extracted 100,000 using a reservoir sampling method without replacement to reduce computational time. Next, these sequences were clustered using the UCLUST method (Edgar, 2010) and mapped by homology using SOAP2 against the 16S rRNA Greengenes v13.5 database (McDonald *et al.*, 2012) and only best hits were retained for further analysis. An abundance raw-count table was built for the seven taxonomical ranks, from phylum to species levels for all samples. In the tables we removed all singleton elements (those appearing just once in a sample) to avoid false positive assignments and then sorted all elements in descending order on the basis of their abundance average using awk and shell scripts.

3.1.6 Synthetic mock communities for validation

To evaluate MetaTrans predictive accuracy for functional analysis, two synthetic mock communities with different expression levels were constituted. Five most abundant microbial genomes were selected based on Qin et al. (Qin *et al.*, 2010) and were downloaded from the NCBI database: *Bacteroides vulgatus* ATCC 8482, *Ruminococcus torques* L2− 14, *Faecalibacterium prausnitzii* SL3/3, *Bacteroides thetaiotaomicron* VPI-5482, *Parabacteroides distasonis* ATCC 8503. A subsample of 1000 genes from each of these microorganisms was selected randomly without replacement to generate a synthetic mock community (4943 reads or 5 Mbp). This mock community was then injected into the Polyester tool (Frazee *et al.*, 2015) to simulate two groups of samples with differential expression level; with each group containing 50 simulated samples as follow: a different expression level has been simulated in one of the two groups, such that 20% of the genes presented a 4-fold overexpression and 20% a 4-fold underexpression.

To test the accuracy of our pipeline for taxonomic assignment, we used one of the 16S rDNA synthetic mock communities provided by the study of Jeraldo et al. (Jeraldo *et al.*, 2014) that resembles an ecological sample in terms of composition and abundance. From this original dataset, we used 2500 unique organisms (14800 reads or 21 Mbp) to simulate differential expression with two replicates of 25 samples each, using Polyester. As for the functional simulation, a different expression level was applied in one of the two groups, such that 20% of the genes presented a 4-fold overexpression and 20% a 4-fold underexpression.

Polyester produced then an output of two groups of samples with a different expression level. To simulate reads with quality scores, we used the ART simulator (Huang *et al.*, 2012) to produce an equal number of reads in FASTQ format to those produced by Polyester. ART was initially trained with our 8 total RNA samples sequenced in a Hi-Seq 2000 Illumina to obtain a quality error model. After simulating FASTQ files we then extracted the quality data and bound it to the FASTA files generating new FASTQ files.

A total of 100 samples for the functional simulation and 50 samples for the taxonomic simulation were then loaded and processed in MetaTrans. To prevent overestimates of accuracy based solely on well-known genomes, we removed from the MetaHIT database those reads that had more than 90% identity with the MetaHIT genes.

To construct ROC curves, we first computed a score (1 – nominal-p-value) for each gene, which allowed us to rank the genes in order of significance or evidence for differential expression between two groups. The score was two sided, that is, it was not affected by the direction of differential expression between the two conditions. Given a threshold value for such a score, we called all genes with scores exceeding the threshold DE (differentially expressed), and correspondingly all genes with scores below the threshold were called non-DE (non-differentially expressed). Considering the genes that were simulated to be DE as the true positive group and the remaining genes as the true negative group, we computed the false positive rate and the true positive rate for all possible score thresholds and constructed a ROC (Receiver Operating Characteristic) curve for the method.

3.2 IBD analysis

3.2.1 Ethics statement

All the experimental protocols were submitted and approved by the local Ethical Committee of the University Hospital Vall d'Hebron (Barcelona, Spain). All volunteers received information concerning their participation in the study and gave a written informed consent.

3.2.2 Design and samples collection protocol

We selected a subset of subjects from a Spanish cohort that were enrolled in a previous study (Pascal *et al.*, 2017). Given that the focus of the study was on finding microbiome differences between the two phenotypes of IBD and healthy subjects, we included in this study 14 CD patients and 14 healthy, 12 of which were first-degree relatives (siblings, children or parents), and 14 UC patients and 14 healthy, 12 of which were healthy first-degree relatives.

Inclusion criteria for patients included: confirmed diagnosis (by endoscopy and histology in the past), clinical remission (for at least 3 months; defined by the colitis activity index (CAI) for UC and by the CD activity index (CDAI) scores (Best *et al.*, 1976), stable maintenance therapy (either amino-salicylates, azathioprine or no drug) and previous history of at least 3 clinical recurrences in the past 5 years. Clinical recurrence was defined by a value of 4 or higher for CAI and higher than 150 for CDAI. Healthy controls (HC, also referenced as healthy relatives, HR) were included without previous history of chronic disease. At inclusion and during the follow-up (every 3 months), diagnostic data was collected, location and behavior of CD, extension of UC, and clinical data including tobacco use and medical treatment.

Exclusion criteria included pregnancy or breast-feeding, severe concomitant disease involving the liver, heart, lungs or kidneys, and treatment with antibiotics during the previous 4 weeks.

Patients with CD and UC who showed recurrence during the study also provided a stool sample at the time of recurrence. For UC, relapse was defined by clinical scores and calprotectin (fecal marker of inflammation). For CD, recurrence was defined by endoscopic criteria and by calprotectin. Antibiotic therapy for the previous 3 months was excluded. Patients and controls were asked to stop any drug intake for 1 week before sampling.

To evaluate differences between relapse and remission, fecal samples from patients were collected in two time-points, at baseline and at one year follow up unless patients underwent a relapse, in that case the sample was collected at the time very close to the beginning of the relapse state.

To assess variability, fecal samples from their healthy relatives were collected also at two time-points, at baseline and after 3 months. Fecal samples were frozen at -20ºC at volunteers' home freezer immediately after collection and then as soon as possible at -80ºC at the laboratory before analysis.

3.2.3 Genomic RNA extraction

Fecal samples were processed for total RNA extraction as described earlier in the section "Genomic RNA extraction" from the "Pilot study" chapter.

3.2.4 rRNA removal and cDNA synthesis and sequencing

Total RNA of one hundred and eleven samples were subjected to an rRNA removal procedure using the Ribo-zero Magnetic kit according to the manufacturer's instruction TruSeq® Stranded Total RNA-Seq Library Preparation Kit from Illumina® (dUTP based, where Read2 dictates strandedness). The samples were then subjected to fragmentation of the remaining RNA molecules; after, complementary DNA (cDNA) of the RNA was synthesized following the same library preparation kit protocol. Each library was sequenced as paired-end 101-bp reads on the Illumina HiSeq 2000 platform (Centre Nacional d'Anàlisi

Genòmica, CNAG, Barcelona, Spain) and produced 222 files (111 paired-end files), which generated a total of 592 Gbp (5.86 Billion reads).

The CNAG sequencing facilities produced, additionally, a total of six technical duplicates of samples which did not reach a certain threshold. From those duplicates, only those with higher base qualities were kept.

3.2.5 Bioinformatic analysis

The analysis was performed using the same computational resources as for the development of the pipeline in the “Pilot study” chapter section “Bioinformatic analysis”.

The Illumina platform provided paired-end reads in a FASTQ format (CASAVA 1.8, Phred + 33) separated into distinct files for each single-end read and for each sample. The microbiome analysis of the data was carried out using a modified version of the previously developed “MetaTrans 1.0” pipeline (Martinez et al., 2016) described in “Pilot study” chapter. We further carried out a gene counts analysis following the same procedure described in Le Chatelier et al (Le Chatelier *et al.*, 2013), who showed that a lower number of gene count was associated with obesity.

3.2.5.1 Pipeline modifications for the IBD study

The metatranscriptomic pipeline, namely MetaTrans, that was initially developed in “Pilot study” chapter, was modified to incorporate some updates that were required to improve its functionality, performance, and annotation in the analysis of IBD samples. Additionally, some bugs were also fixed.

The clustering program USEARCH (v5.2.236) (Edgar, 2010), via its UCLUST algorithm, was recommended to be used in the taxonomical analysis. However, this program, in its free version had had inherent limitations (i.e. RAM memory was limited to few GB). This limitation can be acceptable for small number of samples, but to process the 111 samples of the IBD project, it was not sufficient. An alternative clustering program like CD-HIT (Fu *et al.*, 2012) was tested, but the program lasted up to many hours per sample and constituted a bottleneck in the pipeline. Finally, we decided to use the paid-version of USEARCH (v8.1.1861_i86linux64), which outperformed by many folds any previous clustering tool we used (see https://drive5.com/usearch/cdhit_versions.html for more details) .

The ability to perform differential expression analysis at the 16S rRNA taxonomical level was implemented and incorporated in the pipeline.

The normalization process performed by the DESEq2 package (Love, Huber and Anders, 2014), based on RLE (Relative Log Expression) (Abbas-Aghababazadeh, Li and Fridley, 2018), allows the comparison of features (genes, transcripts,…) between-samples but not within-samples. As recommended by the authors of the package, we scaled by library depth to perform comparisons within samples (to account for different number of reads sequenced per sample) and feature length (e.g. longer genes will tend to have more sequenced reads). The most robust metric accounting for these variations is TPM (Transcripts Per Million) (Wagner, Kin and Lynch, 2012) which represents the number of transcripts seen per million for a specific feature in a certain sample.

At this point, it is important to remark that a proper accurate normalization of metatranscriptomic data, for each sample should be done considering its taxonomical composition obtained from the metagenomic data corresponding to the same sample. Such normalization is left for a future study combining Metagenomic and Metatranscriptomic data (Franzosa *et al.*, 2014, 2018).

The MetaHIT-2014 gene catalog, also known as Integrated Gene Catalog (IGC), was annotated using the non-supervised, i.e. not manually curated, orthologous genes database EggNOG version v3.0, which was launched on November 2011 (Powell *et al.*, 2012). A newer release v4.5 (Huerta-Cepas *et al.*, 2016) was launched on 2015 and accounts for a more accurate annotation and higher coverage of orthologous gene families, along with the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways member annotations. To update IGC-EggNOG annotations we used a script called "eggnog-mapper" to re-map all IGC proteins to the new database. From a total annotation of roughly 40% of the IGC catalog with EggNOGv3.0, this step allowed a partial update of 66% of the EggNOGv3.0 annotated IDs to EggNOGv4.5.

Aside from the EggNOG functional annotation in the IGC gene catalog, the genes were also annotated using the last free release of the manually curated orthologous groups database KEGG Orthology v.59 (Kanehisa and Goto, 2000; Kanehisa *et al.*, 2017). Nevertheless, the functional profiling for this database had not yet been implemented. The pipeline is now able to perform abundance profiles of KEGG orthologs at the four levels of the KEGG functional hierarchy.

For the visualization of the metabolic pathways expressed or activated in a certain condition we used the iPath2 explorer (Letunic *et al.*, 2008; Yamada *et al.*, 2011) to produce visual maps of the metabolic pathways activity. This tool was introduced in the MetaTrans pipeline programmatically over HTTP to enable metabolic pathway plots when necessary. Recently, a new update of the tool, iPath3 (Darzi *et al.*, 2018), was released as of 2018 and we made the appropriate pipeline modifications to include it. One of the new features included allows annotation of EggNOG orthologous IDs v4.5.

Statistical analyses were improved by adding significance tests in the PCA (principal component analysis) distance matrices by means of the Permutational Multivariate Analysis of Variance (Adonis Test using the Vegan R package

(Oksanen *et al.*, 2018)) test using the wrapper QIIME (Caporaso *et al.*, 2010) python script “compare_categories.py”.

Along the analysis of the differential expressed features (a.k.a. genes, transcripts, orthologous IDs, …) we also introduced several changes to expand statistical data and visualization of results using the R package DESeq2 (Love, Huber and Anders, 2014). In the analysis of the differences among groups of samples, i.e. between-samples or beta-diversity analysis, we added dendrogram plots using distance matrices of samples to represent in a tree their similarities. Several distance metrics like Euclidean, which considers abundances and Bray-Curtis, which accounts for composition as well, were used as input. Further, other hierarchical clustering linking methods were also explored like “upgma”, “complete”(default), and “ward.d2” (Singh, 2008) in the clustering analyses. The alpha-diversity, i.e. within-sample evenness, analysis was also extended by the inclusion of other diversity indices (Chao2, Pielou, Simpson) to check for different properties to assess evenness, as there is no clear consensus on which is most necessary (Morgan *et al.*, 2012; Franzosa *et al.*, 2014; Kvalseth, 2015; Oksanen, 2015; Ricotta, 2017).

To increase the power of significance tests in the differential expression analysis (DEA) we incorporated a pre-filtering step to keep only rows having at least ten reads in total. As explained by the author, this allows reducing the memory, and increases the speed of the transformation and testing functions.

3.2.6 Databases used in this study

For the functional and taxonomical analysis, we performed comparisons using different database annotations. In the case of the functional database we selected the two types of annotations available in the integrated MetaHIT14 catalog of microbial genes from the human gut (Li *et al.*, 2014), i.e. the curated orthologs genes annotation KEGG Orthology v56 (KO) database (Kanehisa and Goto, 2000; Kanehisa *et al.*, 2017, 2019) and the non-curated and automated annotation EggNOG v3.0 database (Powell *et al.*, 2012; Huerta-Cepas *et al.*, 2016). The EggNOG database provides two types of annotation levels, at protein family level, or COG functional categories (Tatusov *et al.*, 2001), and at orthologous genes level. The KEGG database provides five functional annotation levels, four use the KEGG BRITE functional hierarchies (first level more generic, last level being more specific), and the last one at KEGG orthologous genes level. On the other hand, the taxonomic annotation of genes was performed using the Greengenes v135 database ranks (DeSantis *et al.*, 2006). We used the seven taxonomical ranks available for the analyses (i.e. Phylum, Class, Order, Family, Genus, and Species).

3.2.6.1 Statistics

Each annotation level was measured as Euclidean distances matrices, necessary for downstream analyses. The statistical test used to find differences between groups was Adonis (Permutational Multivariate Analysis of Variance or PERMANOVA), from the Vegan R package (Oksanen *et al.*, 2018), using the QIIME (Caporaso *et al.*, 2010)“compare_categories.py” wrapper script.

To measure alpha-diversity, we employed three of the most used indexes: one based on diversity (Shannon index), another to measure evenness (Pielou’s Index), and one to measure richness (Chao1 index). Hypothesis testing was conducted by means of the non-parametric Mann-Whitney U test.

To perform univariate or bivariate analysis we used the “ggstatsplot” and “ggplot2” R packages (Wickham, 2016; Patil and Powell, 2018) within the Jamovi statistical platform (R Core Team, 2018; The jamovi Project, 2019). The “ggstatsplot” package allows us performing the most suitable statistical test according to the nature of the groups being compared (see Figure 3.2).

The analysis of the differential expressed genes (DEG) was performed using a the “DESeq2” R package (Love, Huber and Anders, 2014), which estimate variance-mean dependence in raw gene/functions count data from RNA sequencing, and base their differential expression test on a model using the negative binomial distribution. This package was added as part of the MetaTrans pipeline developed in the first part of the dissertation.

Dendrograms were used to perform hierarchical cluster analysis on samples based on sample to sample distances obtained from calculations on relative abundances of samples. We used the regularized log transformation of counts as suggested by the author of the package “DESeq2” (Love, Huber and Anders, 2014) used for the detection of differentially expressed features (taxa or functions), which minimizes the influence of “low counts”. The data matrix was then used as input using the “pvclust” function (Suzuki and Shimodaira, 2015) adapted to use Bray-Curtis dissimilarity distances (script by Niel Shanson, https://git.io/JeRRD). The recommended agglomeration method used for the hierarchical clustering was “Ward.D2”, which is known to be effective in this type of data. This class of dendrograms uses an improved calculation of p-values based on multiscale bootstrap resampling (called Approximately Unbiased, or AU), which outperforms p-values calculates using normal bootstrap resampling. In absence of a statistical test for clustering analysis, though recently a new package is on development to address this issue https://git.io/Je0bY, we selected the best dendrogram based on their best AU p-values. This type of p-values ranges from 0 to 1, being 1 the value indicating that a particular cluster is strongly supported by the data.

For the identification of KEGG metabolic pathways significantly over or under represented from differentially expressed genes we applied a Fisher’s exact test implemented in a perl script within the FMAP tool (Kim *et al.*, 2016).

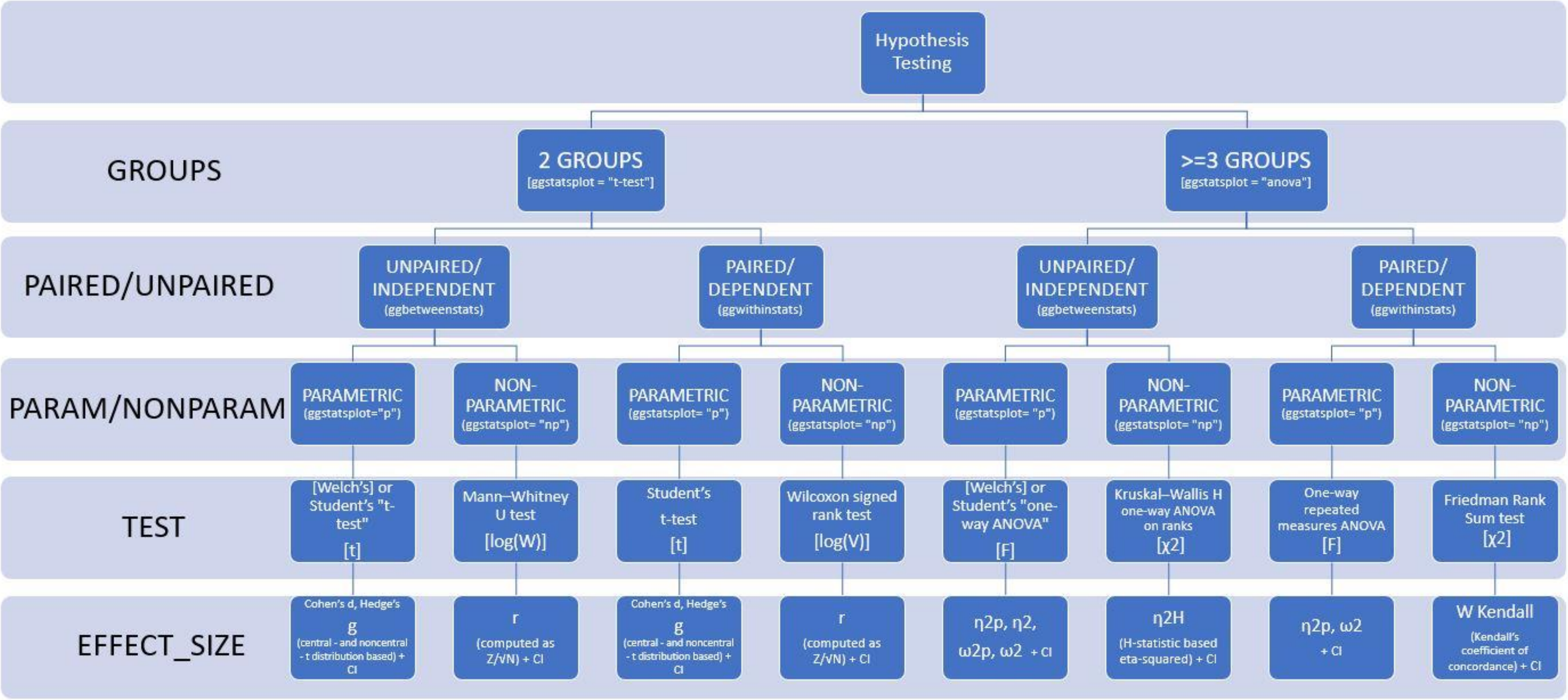

Hypothesis Testing
GROUPS
2 GROUPS [ggstatsplot = "t-test"]
>=3 GROUPS [ggstatsplot = "anova"]
PAIRED/UNPAIRED
UNPAIRED/ INDEPENDENT (ggbetweenstats)
PAIRED/ DEPENDENT (ggwithinstats)
UNPAIRED/ INDEPENDENT (ggbetweenstats)
PAIRED/ DEPENDENT (ggwithinstats)
PARAM/NONPARAM
PARAMETRIC (ggstatsplot="p")
NON-PARAMETRIC (ggstatsplot= "np")
PARAMETRIC (ggstatsplot= "p")
NON-PARAMETRIC (ggstatsplot= "np")
PARAMETRIC (ggstatsplot= "p")
NON-PARAMETRIC (ggstatsplot= "np")
PARAMETRIC (ggstatsplot= "p")
NON-PARAMETRIC (ggstatsplot= "np")
TEST
[Welch's] or Student's "t-test" [t]
Mann–Whitney U test [log(W)]
Student's t-test [t]
Wilcoxon signed rank test [log(V)]
[Welch's] or Student's "one-way ANOVA" [F]
Kruskal–Wallis H one-way ANOVA on ranks [χ2]
One-way repeated measures ANOVA [F]
Friedman Rank Sum test [χ2]
EFFECT_SIZE
Cohen's d, Hedge's g (central - and noncentral - t distribution based) + CI
r (computed as Z/√N) + CI
Cohen's d, Hedge's g (central - and noncentral - t distribution based) + CI
r (computed as Z/√N) + CI
η2p, η2, ω2p, ω2 + CI
η2H (H-statistic based eta-squared) + CI
η2p, ω2 + CI
W Kendall (Kendall's coefficient of concordance) + CI

Figure 3.2 Available statistical tests in the "ggstatsplot" R package.

Chapter 4.

Results

4 Results

4.1 Pilot study

The Illumina sequencer machines produced an average of 22 million paired-end reads of short-length (76 bp) per sample which were mapped against functional databases. We also compared two methods of taxonomic analysis; one using 16S rDNA V4 amplicons and the other 16S rRNA extracted from total RNA.

4.1.1 Pipeline validation

As described in Figure 3.1, the pipeline, consisting of four major steps (filtering, sorting, and functional and taxonomic analyses), included tools implemented with multi-threading options and used the most updated functional human gut database (MetaHIT-2014)(Li *et al.*, 2014).

In order to validate MetaTrans in terms of taxonomic analysis, we compared different available methods such as 1) 16S rRNA sequences analyzed with the SOAP2 tool (Li *et al.*, 2009) and the Greengenes database (McDonald *et al.*, 2012); 2) total RNA analyzed with MG-RAST (Wilke *et al.*, 2015); 3) mRNA sequences analyzed with the Kraken tool (Wood and Salzberg, 2014); 4) mRNA analyzed with SOAP2 and the MetaHIT-2014 database. Kraken is a taxonomic sequence classifier that assigns taxonomic labels to short DNA reads. To classify a sequence, Kraken maps each k-mer in the sequence to the lowest common ancestor (LCA) of the genomes that contain that k-mer in a database (NCBI-bacterial/archaeal genomes). For rRNA identification, MG-RAST uses a BLAT similarity search for the longest cluster representative against the M5rna database that includes SILVA, Greengenes and RDP databases. To compare the different methods, we used all the sequences of one of our processed fecal sample (#1_BF), for which, we generated 39 million paired-end reads. After processing the reads by MetaTrans, we recovered 1.6 million of 16S rRNA paired-end sequences and 700000 mRNA hit against the MetaHIT-2014 database that were then used for the methods comparison. As shown in the

Figure 4.1, 16S rRNA sequences analyzed with the SOAP2 tool and the Greengenes database (rRNA.GG.SOAP2) presented at the phylum level very similar results with those of the MG-RAST server and displayed very low proportion of unclassified reads (< 5%). mRNA analyzed with Kraken and the NCBI bacterial database showed also low proportion of unclassified reads (1%) but higher relative abundance of Euryarchaeota than the two previous methods, which could be due to the higher copy number of 16S rRNA gene found in Bacteria compared to Archeae (Lee *et al.*, 2009). Only mRNA analyzed with SOAP2 and the MetaHIT-2014 database (mRNA.MetaHIT) presented a very high percentage of unclassified reads.

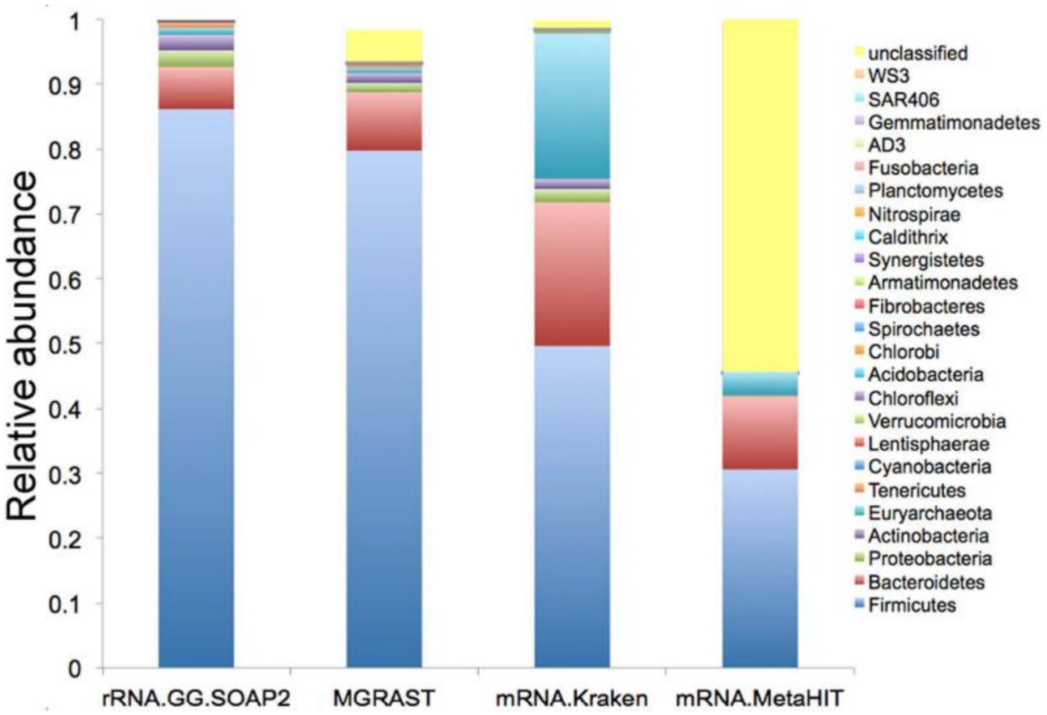


Figure 4.1 Comparison of taxonomic classification methods.
Taxonomic assignment in terms of abundance for the fecal sample #1_BF using 16S rRNA sequences mapped with SOAP2 against Greengenes (rRNA. GG.SOAP2), the whole sample analyzed with MG-RAST (MG-RAST), mRNA assigned with Kraken (mRNA. Kraken) and mRNA mapped with SOAP2 against MetaHIT-2014 (mRNA.MetaHIT). This bar plot shows similar taxonomic profiles between rRNA.GG.SOAP2 and MG-RAST whereas they differ with mRNA.MetaHIT and mRNA.Kraken. Unclassified reads are more abundant in the last method.

To assess the accuracy of MetaTrans for taxonomic profiling, we also constructed two synthetic mock communities of 25 samples each. We applied a differential expression such that 20% of the genes presented a 4-fold over- expression and

20% a 4-fold underexpression between the two communities and the sensitivity and specificity of MetaTrans were evaluated using a receiver operating characteristic (ROC) curve (see Methods section; Figure 4.2-A). We obtained an AUC (area under the curve) of 0.704, which showed a fair accuracy of the method.

In order to validate MetaTrans in terms of functional analysis between two microbial communities, we also constructed two mock communities of 50 samples each and simulated a differential gene expression between the two communities as described above. Each sample contained 1000 genes randomly selected from five microorganisms commonly found in the gut microbiome (see in Methods section). As for the simulation of 16S rRNA dataset, we evaluated the performance of MetaTrans using a ROC curve (Figure 4.2-B). We obtained an AUC of 0.887, which showed a good accuracy of the method. To test our pipeline in terms of functional analysis with real metatranscriptomic data, we recovered and processed part of the dataset published in a previous study (Leimena *et al.*, 2013). This dataset consisted of paired-end reads obtained from the content of a human small intestine sample (42.2 million sequence reads for both ends). For these analyses, our pipeline was adapted to match the reads to the COG database (Clusters of Orthologous Groups, containing about 190,000 annotated functions and 25 categories of functions) using BLASTP, as performed in Leimena et al. We obtained all the 23 functional categories as described in Leimena et al. and in similar proportions (Figure 4.3).

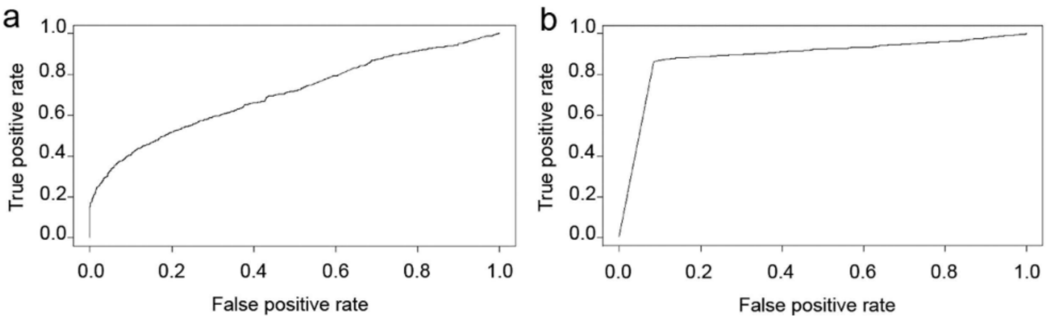


Figure 4.2 Performance of MetaTrans for analyses of mock community simulations
ROC curves of (a) taxonomic and (b) functional mock community simulations, with 50 and 100 samples, respectively. AUC of 0.704 and of 0.887 were obtained for taxonomic and functional simulations, respectively.

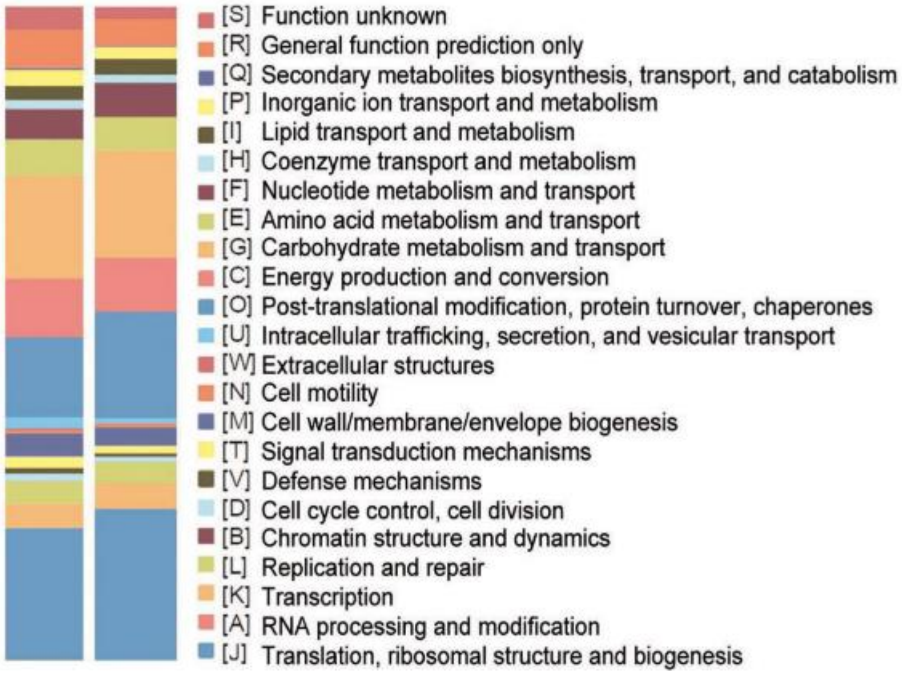


Figure 4.3 Pipeline validation with another study.
In order to test whether our pipeline provided similar results to those obtained using a previously reported tool, we analyzed part of a published dataset (Leimena *et al.*, 2013) using our pipeline. We obtained similar functional categories (left) to those described in Leimena et al. 2013 (right).

4.1.2 Experimental design

To test our pipeline with RNA-seq newly generated, we performed RNA sequencing in two types of experimental designs: “total RNA” and “rRNA removal” experiments. The objective of the “total RNA sequencing” experiment was to recover both the functional and taxonomic profile of each active microbial community in an unbiased manner. We performed this experiment on eight stool samples from four individuals that were collected in a previous study (Manichanh *et al.*, 2013). Moreover, in order to detect functional variations for each participant, samples were collected before and immediately after three days of a flatulogenic diet, as detailed in the Methods section. We then compared the 16S rRNA sequences with the 16S rDNA sequences that we recovered from our previous study (Manichanh *et al.*, 2013) after PCR amplification of extracted genomic DNA of the same samples. We envisaged that this comparison would indicate whether the microbes detected by the 16S rDNA gene survey were also those functionally active. The objective of the “rRNA removal” experiment was to test how the rRNA depletion step would increase the recovery of number of expressed genes. This experiment was performed on four additional stool samples obtained from four individuals.

As paired-end reads have been shown to recover fewer false positives than single ones (González and Joly, 2013), we assembled, when possible, the single end reads using the Fastq-Join program before performing gene prediction by FragGenScan (Figure 3.1).

4.1.3 Data and output descriptions

The two experiments generated the following datasets: 318 million paired-end reads (76 bp) generated from the “total RNA” experiment (about 20 million paired-reads per sample; Table 3) and 219 million paired-end reads (76 bp) generated from the “rRNA removal” experiment (about 27 million paired-reads per sample; Table 4). For the “total RNA” experiment, we recovered an average of 78% high quality reads, 74% of rRNA/tRNA and 4.3% of non-rRNA/tRNA (e.g. potential mRNA), as expected. For the “rRNA removal” experiment, we obtained 55% of high-quality reads, 2.7% of rRNA/ tRNA and 52.3% of potential mRNA. As expected, the proportion of potential mRNA recovered from “rRNA removal” experiment was 10 fold higher than in the “total RNA” experiment. However, the median number of unique orthologous IDs was only 1.27-fold higher in the “mRNA removal” experiment (11541 versus 9032). Furthermore, we observed that the overlapping step allowed recovery of a longer read length for 42% of the non-rRNA/tRNA reads for the two experiments.

Table 3 Description of the outputs from each analysis step of the "total RNA" experiment.

	#1_BF Diet	**#1_AF Diet**	**#2_BF Diet**	**#2_AF Diet**	**#3_BF Diet**	**#3_AF Diet**	**#4_BF Diet**	**#4_AF Diet**	**Average**
Raw reads (X2)	39534526	44137900	36621522	35376906	44550458	46221230	29230252	43061504	39841788
After quality control (%)	80	78	79	75	77	76	78	82	78.1
rRNA/tRNA (%)	75	73	76	72	76	71	73	77	74.1
Non rRNA/tRNA (%)	5.7	5.2	3.0	3.1	1.4	5.0	5.8	5.3	4.3
After paired-end overlapping (%)	4.7	4.2	2.4	2.5	1.2	4.4	4.7	4.5	3.6
After FragGeneScan (%)	3.9	3.4	1.8	2.0	1.0	3.5	4.0	3.8	2.9
After CD-HIT (%)	2.7	2.3	1.0	1.3	0.6	2.6	2.8	2.4	2.0
MetaHIt ids	965435	906270	350852	438399	235689	1072589	805893	1077089	731527
Unique MetaHit ids	292371	288398	111463	140293	85343	278887	201651	252232	206330

Table 4 Description of the outputs from each analysis step of the "rRNA removal" experiment.

	#5	**#6**	**#7**	**#8**	**Average**
Raw reads (X2)	50948448	51008628	61034072	55883372	54718630
After quality control (%)	40.6	67.1	62.3	49.6	54.9
rRNA/tRNA (%)	3.7	4.8	0.7	1.4	2.7
Non rRNA/tRNA (%)	36.9	62.3	61.7	48.3	52.3
After paired-end overlapping (%)	26.8	45	44.3	35.5	37.9
After FragGeneScan (%)	23.8	40.9	41.6	32.5	34.7
After CD-HIT (%)	4	7.2	6.7	4.6	5.6
MetaHIt ids	7010023	12337711	16396625	10614486	11589711
Unique MetaHit ids	221132	482156	495872	320431	379898

4.1.4 Computer bottlenecks

Metatranscriptomic as well as metagenomic approaches are computationally very expensive (CPUs and RAM). In order to speed up the analysis, our pipeline was optimized by means of multi-threading software. In order to optimize the runtime, we tested several aligner tools such as DIAMOND-BLASTP (Buchfink, Xie and Huson, 2014), SOAP2 (Li *et al.*, 2009) and BLASTP (Edgar, 2010) to map one of our dataset (#1_BF) against the MetaHIT-2014 database, containing human gut microbiome genes. The three tools provided very similar number of matched eggNOG IDs (Figure 4.4). However, SOAP2 and DIAMOND-BLASTP were 6600 and 480 fold much faster than BLASTP, respectively. We finally implemented SOAP2 and DIAMOND-BLASTP in our pipeline. The bottleneck still remains in the first steps of the analysis, in particular for the sorting and clustering steps. Therefore, to be able to perform these analyses in a reasonable timeframe, we recommend a minimum of 10 CPUs and 16 GB of RAM (size of the database or the query to be loaded). As an example, to analyze a sample, for which 39 million paired-end reads were generated and about 1 million of potential genes were sorted out, 2 hours and 21 min was required with the following settings: 10 CPUs and 16 GB of RAM. The cost of a current computer with these features could approximate 3000 dollars.

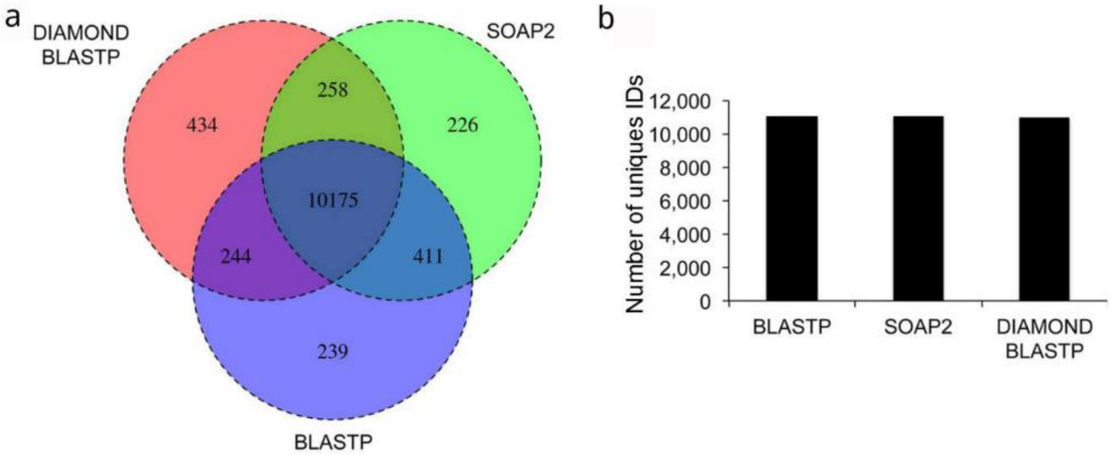


Figure 4.4 Mapping comparisons between short-read aligners.
Similarity in functional mapping between BLASTP, DIAMOND-BLASTP and SOAP2 against the MetaHIT-2014 database using dataset from sample #1_BF as shown by a Venn diagram (a) and the plot of the total number of unique IDs that have a match against the MetaHIT-2014 database (b).

4.1.5 Taxonomic analysis

To describe the active microbial composition of our stool samples from the "total RNA" experiment, we mapped the reads labeled as rRNA/tRNA against the Greengenes (v13.5) 16S rRNA database. To speed up the taxonomic analysis, we randomly selected a reasonably high number of reads, namely 100000, a much higher number than most studies performing 16S rRNA gene surveys. At the phylum, family, genus and species levels, we identified 7, 29, 49 and 70 groups of microbes, respectively, with at least 1% of sequences in at least one sample, in order to avoid false positives. Four phyla accounted for 99.3% of the dataset: Firmicutes (87%), Bacteroidetes (8.1%), Actinobacteria (1.9%), and Proteobacteria (1.8%). At the family level, Lachnospiraceae (52.2%), Ruminococcaceae (18%), unknown Clostridiales (11%), Bacteroidaceae (5%), Erysipelotrichaceae (1.9%), Clostridiaceae (1.9%), and Porphyromonadaceae (1.1%) accounted for 91% of the total relative abundance. Comparative 16S rDNA and 16S rRNA sequence analysis indicated significant differences (q-value < 0.05, Kruskal-Wallis) between relative mean abundance of the 16S genes detected at all phylogenetic levels from phylum to species, suggesting that the 16S rDNA survey did not provide the profile of the active microbial community. Indeed, at RNA level, Firmicutes might be a more dominant part of the metabolically active bacteria than suggested at DNA level (average of 87% in rRNA vs. 53% in rDNA sequence libraries) (Figure 4.5-A). At the family level, Lachnospiraceae (52% for rRNA vs. 26% for rDNA) was a significantly more active component than Bacteroidaceae (5.2% for rRNA vs. 28% for rDNA). At the genus level, an unknown Lachnospiraceae, *Blautia* (a Lachnospiraceae genus) and an unknown Clostridiales predominated the rRNA libraries, with a total mean of 50%. In contrast, in the rDNA libraries, *Bacteroides*, an unknown Ruminococcaceae and an unknown Lachnospiraceae totaled 50%. Interestingly, most 16S rDNA surveys and metagenomic approaches previously proposed Bacteroidaceae as a major actor in gut function and revealed Lachnospiraceae as the most active group of microbes (Human Microbiome Project). Indeed, members of the Lachnospiraceae family have been linked to obesity and protection against colon cancer in humans. This protective function is mainly due to the association of many species within the group with the production of butyric acid that is important for both microbial and host epithelial cell growth (Meehan and Beiko, 2014).

To compare the number of taxa present at DNA and RNA levels, we first normalized the number of sequence reads per library to 1,952 and used the Friedman test. We observed that the flatulogenic diet caused an increase in Bifidobacteriaceae and more specifically *Bifidobacterium longum* ($P < 0.05$), at the RNA level (Figure 4.5-B) but not at the DNA level. As Bifidobacteriaceae is well-known as a saccharolytic bacterial group, this result would be consistent with a consumption of a flatulogenic diet.

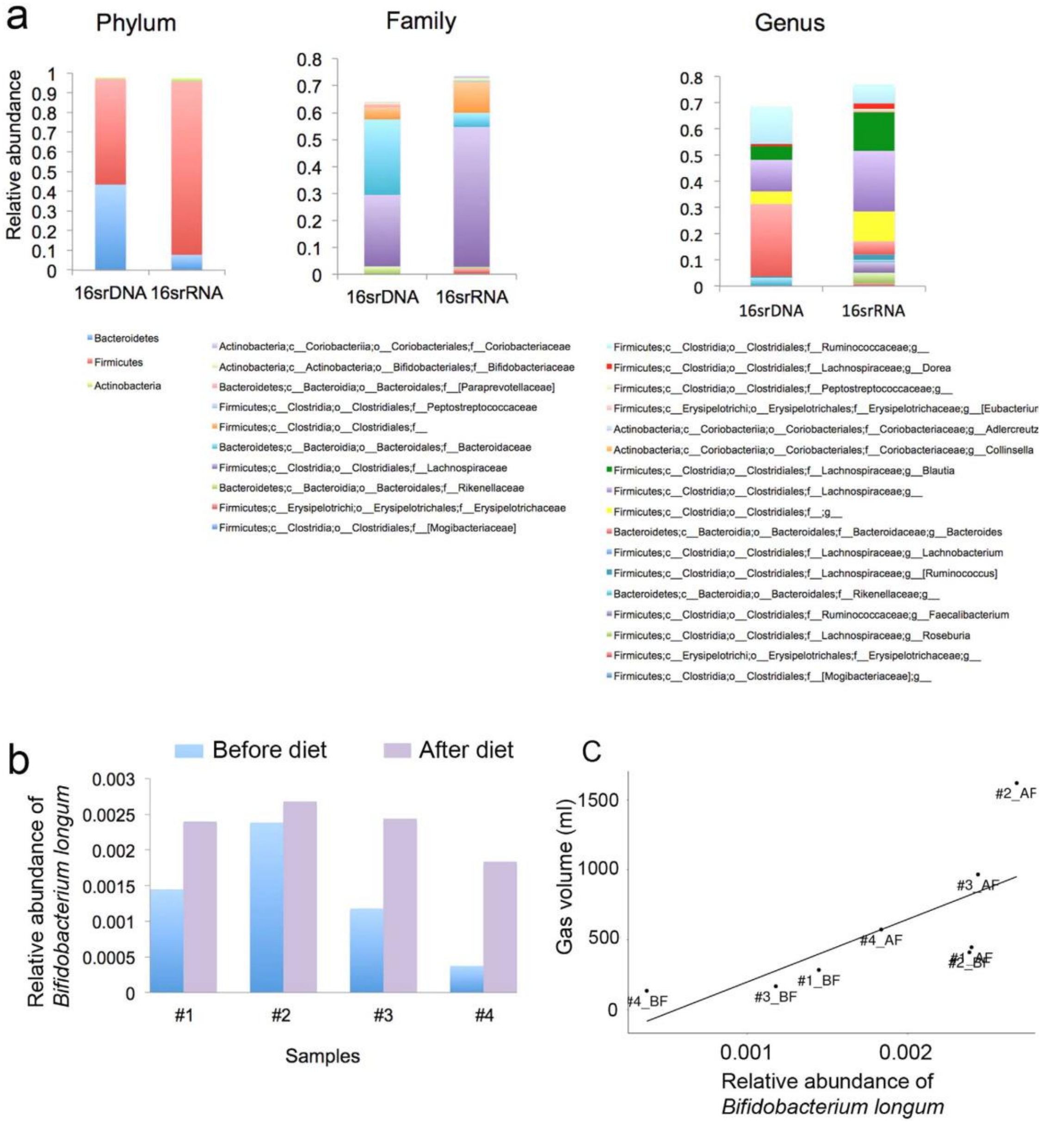


Figure 4.5 Taxonomic analysis at the DNA and RNA levels
(a) Significant differences between relative mean abundance of the 16S rRNA and 16S rDNA libraries at the phylum, family and genus levels (q-value < 0.05). (b) Effect of diet at the RNA level on the increase in relative abundance of *Bifidobacterium longum* (P < 0.05). (c) Correlation between volume of gas and relative abundance of *Bifidobacterium longum* (r = 0.92; P = 0.002; Spearman)

To assess the link between a flatulogenic diet and intestinal gas production, we correlated the microbiome composition and functions with the volume of gas produced by the subjects and measured before and after the flatulogenic diet. The volume of intestinal gas was found, at the DNA level, significantly and positively correlated with *Blautia* (r = 0.83; P = 0.01), a genus belonging to the Firmicutes phylum. Interestingly, several species belonging to this genus such as *Blautia hydrogenotrophica*, are capable of metabolizing H2/CO2 to acetate (Bernalier *et al.*, 1996). At the RNA level, only *Bifidobacterium longum* was positively correlated with the volume of gas (r = 0.92; P = 0.002; Figure 4.5-C). At the level of categories of functions, we observed that the volume of intestinal gas was significantly and positively correlated with two functional categories: “Inorganic ion transport and metabolism” and “Extracellular structures”; and negatively correlated with one functional category: “Cell motility”. Ninety-one orthologous IDs such as those involved in amino acids metabolism presented a significant positive correlation with the volume of gas, meanwhile 14 orthologous IDs such as those involved in energy metabolism were negatively correlated.

4.1.6 Functional analysis

To characterize the active microbial functions of the eight stool samples from the “total RNA” experiment, reads labeled as non-rRNA/tRNA were subjected to FragGeneScan to predict putative genes and were then mapped against a known protein database, namely MetaHIT-2014 (Integrated Gene Catalog from human gut microbiome) (Li *et al.*, 2014). The MetaHIT-2014 database was annotated following the Kyoto Encyclopedia of Genes and Genomes (KEGG) and the evolutionary genealogy of genes non-supervised orthologous group (eggNOG) databases and it contains 9.9 million non-redundant genes identified in the human gut. Mapping against MetaHIT-2014 allowed us to assign an average of 1.85% (731,527 of reads on average) of the high-quality reads to an average of 206,330 non-redundant MetaHIT IDs or genes (ranging from 85,343 to 288,398) and 25 clusters of orthologous groups (COGs) or functional categories per subject. Table A 5 shows the distribution of the annotated orthologous groups, with carbohydrate transport and metabolism being the most abundant known functional group, as expected for the human gut microbiome (Li *et al.*, 2014).

Further analyses were performed using the DESeq2 package. To detect differentially expressed functions and categories of functions, we applied a Principal Component Analysis to the matrix of abundance count generated after the mapping step. In terms of global eggNOG IDs, the two samples from each subject clustered and were located far from those of the other subjects (PC1 = 34%; Figure 4.6-A), while in terms of functional categories, samples clustered according to the effect of the flatulogenic diet for three of the subjects (PC1 = 64%; Figure 4.6-B). These results suggested that each individual has a specific set of functions and that a flatulogenic diet influenced families of functions.

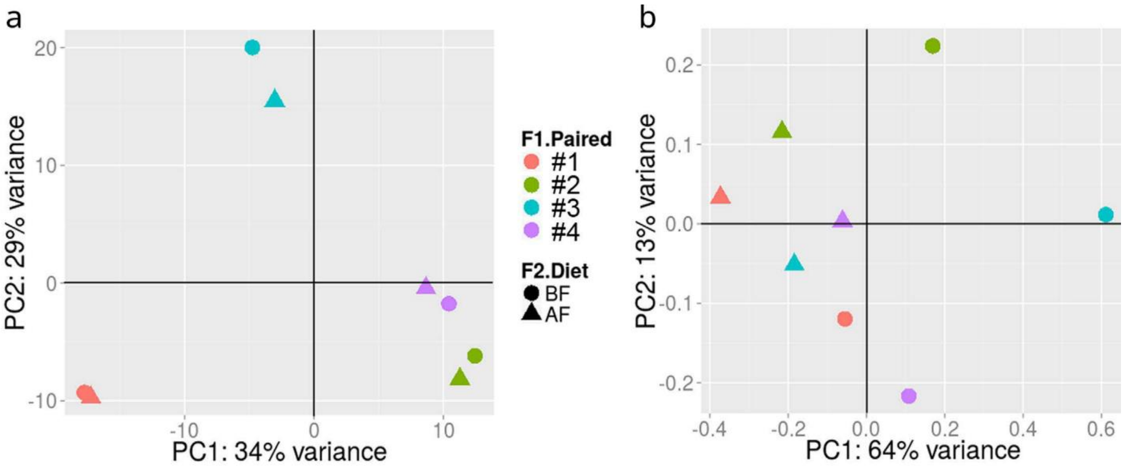


Figure 4.6 PCA analysis of functional databases
Principal Component analysis of the matrix of eggNOG IDs (a) or COG functional categories (b) showed that the two samples, before and after diet, clustered when all functions were taken into account but not when categories of functions were considered

To identify differentially regulated functions or categories of functions, we computed “FoldChange” on a matrix of raw count functions before and after the flatulogenic diet and then tested whether the mean of the log ratios was significantly different from zero following false-discovery rate (FDR) correction (indicating a pattern of up- or down-regulation of functions). As an effect of diet, we observed a significant increase in one category (q-value < 0.01) involved in “Defense mechanisms” and three down-regulated categories involved in “Translation, ribosomal structure and biogenesis”, “Energy production and conversion” and “Carbohydrate transport and metabolism” (Figure 4.7-A). We also identified 27 down-regulated orthologous IDs (with a log2FoldChange < − 1;

q-value < 0.01) (Figure 4.7-B). These were plotted into a network of metabolic pathways using the iPath2 tool (Figure 4.7-C). Among the up-regulated functions, the most abundant was found to be involved in bacterial secretion (Type IV secretory pathway, VirD4 components). The most abundant down-regulated functions were involved in translation (ribosomal protein and GTPases - translation elongation factors), glycolysis (GAPDH - Glyceraldehyde- 3-phosphate dehydrogenase), nucleotide metabolism, vitamin B6 biosynthesis, energy metabolism, and CO dehydrogenase/acetyl-CoA synthase. The latter is central to the acetate production pathway.

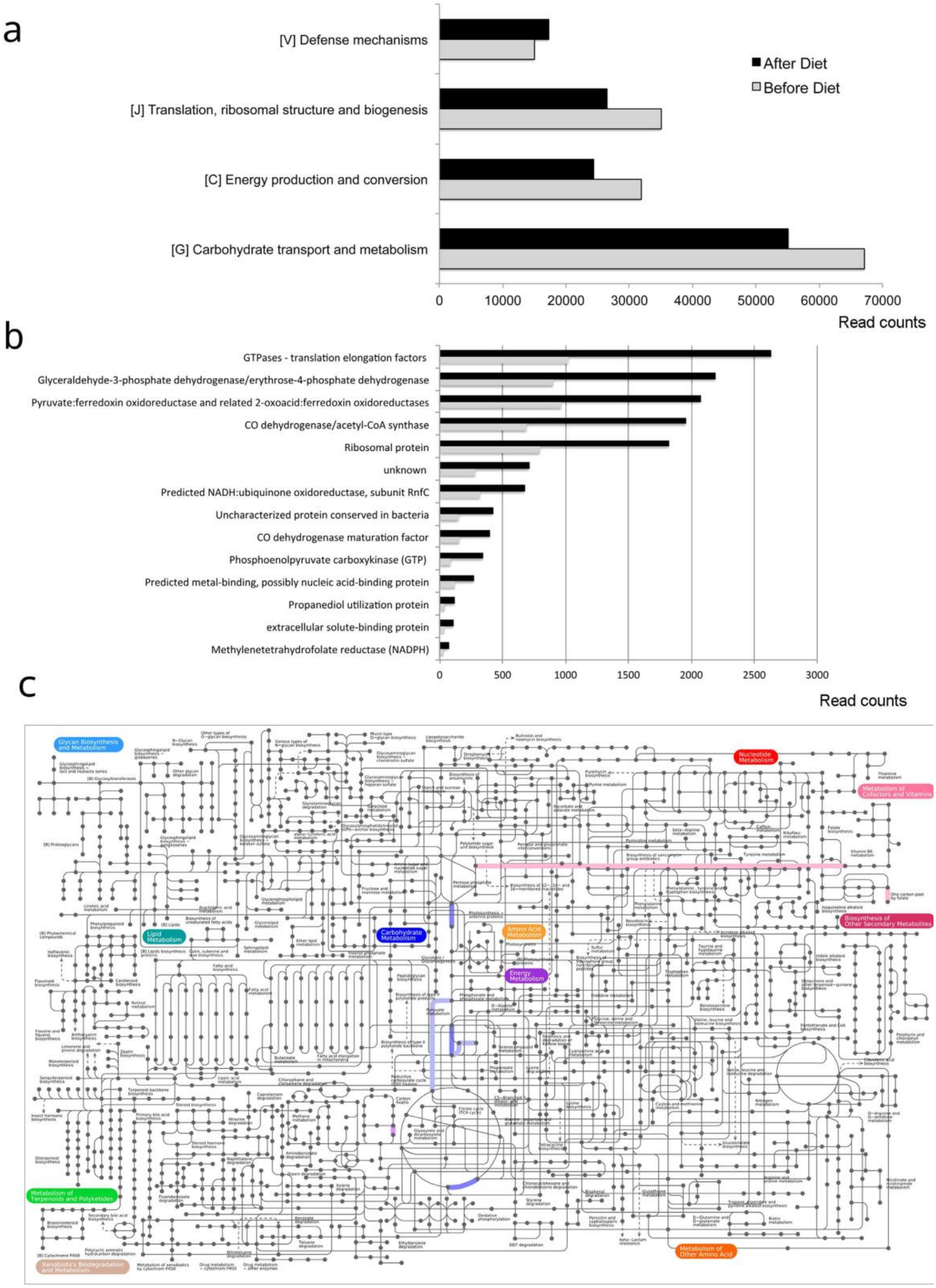


Figure 4.7 Effect of a flatulogenic diet on gene expression

Functional categories (a) that were up and down- regulated and Orthologous IDs (b) that were down-regulated as an effect of the diet challenge (q-value < 0.05). (c) The down-regulated functions plotted into a metabolic pathway network using the iPath2 tool.

4.1.7 Comparison with an existing web application server

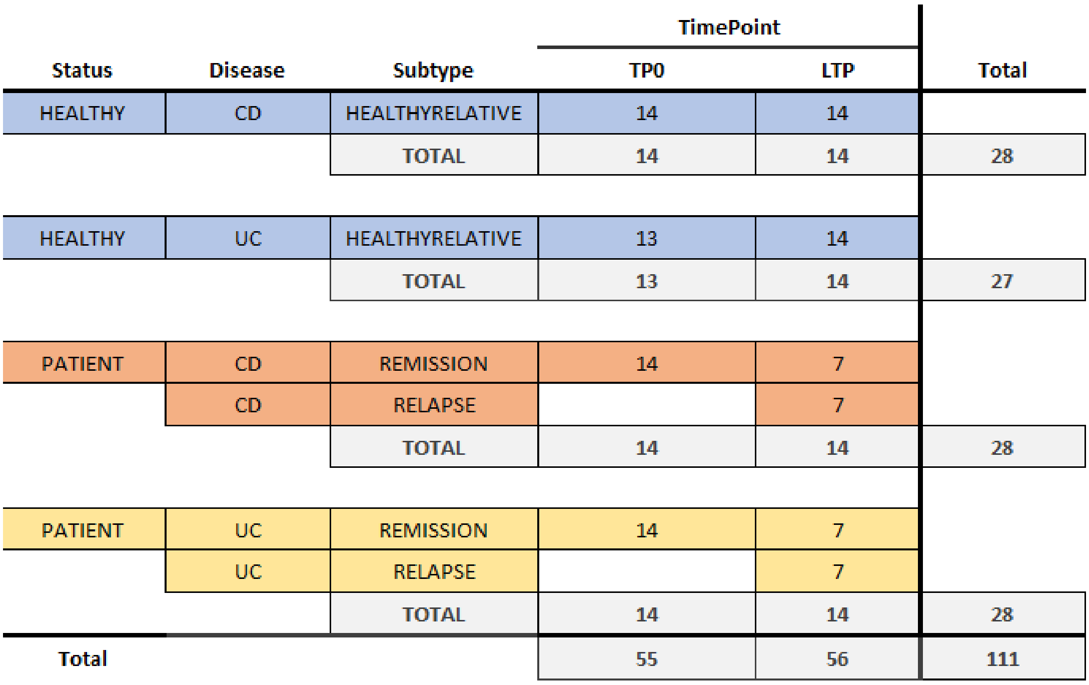
In order to compare the results of our pipeline with those of MG-RAST, one of the few web application servers for metagenomic and metatranscriptomic anal- ysis, we loaded one of our dataset (#1_BF) into the server. The comparison between our pipeline and MG-RAST showed that our pipeline provided, after CD-HIT, a much higher proportion of mapped queries (69% versus 3.2%), probably due to the use of the MetaHIT-2014 database, which only contains genes from the human gut microbiome. This result confirms the necessity to use specific databases. In terms of runtime, after sending three time our dataset for analysis, it took two, three and seven days for MG-RAST to send us back the results, which is much longer than our pipeline (around 2–3 hours). Since MG-RAST is a web application, the time needed to obtain the results depends on several parameters. The Internet connection speed of the users will condition the time needed to upload their dataset. Next, the speed of the analysis will depend on the priority assigned to the project, the size of the dataset and the current server load (as specified by the MG-RAST user manual). Furthermore, the MG-RAST does not provide yet any tools for comparing gene expression levels and we believe that our pipeline would be also more convenient for large metatranscriptomic projects in terms of runtime, providing that the users can handle the analysis through the locally installed pipeline.

4.2 IBD analysis

4.2.1 Study cohort and experimental design

We enrolled 56 subjects in total, 28 IBD patients, of which 14 were UC and 14 were CD, and 28 healthy individuals (or healthy controls, HC), most of which were relatives of patients. All subjects were recruited at a single site in Spain. Adult IBD patients were enrolled in the study with the conditions that they did not take antibiotics for at least two months and were under remission. All them were included in a follow-up study of two time points: basal status (all patients in remission) and final status (all patients in relapse or in remission for one year). However, the sample UC.24.0 (basal timepoint) had to be discarded from the analysis due to the lack of RNA material during the extraction procedure. Thus, a total of 111 fecal samples finally available for microbiome analysis as described in Table 5.

Table 5 Summary of number of samples per health status, and timepoint.



			TimePoint		
Status	**Disease**	**Subtype**	**TP0**	**LTP**	**Total**
HEALTHY	CD	HEALTHYRELATIVE	14	14	
		TOTAL	14	14	28
HEALTHY	UC	HEALTHYRELATIVE	13	14	
		TOTAL	13	14	27
PATIENT	CD	REMISSION	14	7	
	CD	RELAPSE		7	
		TOTAL	14	14	28
PATIENT	UC	REMISSION	14	7	
	UC	RELAPSE		7	
		TOTAL	14	14	28
Total			55	56	111

Most of the CD patients presented the disease in the ileum (21%) and in the ileo-colonic region (64%). This study did not include CD patients with colonic disease location, being less frequent than the other CD subtypes. Characteristics of the CD and UC patients and their healthy relatives are listed in Table 6 and Table 7 respectively.

Table 6 Description of the characteristics of CD patients and their healthy relatives.

Baseline clinical characteristics	**CD (N=14)**	**Healthy relatives of CD (N=14)**
Male/Female (%)	5/9 (35.7/64.3)	7/7 (50/50)
Mean age (SD) at samples collection	32.5(10.5)	48.4 (15.8)
Median BMI (IQR)	20.2 (19.6-25.3)	24.5 (23.61-28.1)
Mean duration of disease (SD) at sampling	7.35 (6.5)	
Disease location (Montreal classification)		

Baseline clinical characteristics	CD (N=14)	Healthy relatives of CD (N=14)
L1 ileal (%)	3 (21.4)	
L2 colonic (%)	0	
L3 ileocolonic (%)	9 (64.3)	
L1 + L4 ileal and isolated upper GIT (%)	1 (7.1)	
L3 + L4 ileocolonic and isolated upper GIT (%)	1 (7.1)	
Disease behaviour at surgery (Montreal classification)		
B1 non-stricturing, non-penetrating (%)	1 (7.1)	
B2 stricturing (%)	8 (57.1)	
B3 penetrating (%)	3 (21.4)	
B1p non-stricturing, non-penetrating and perianal disease (%)	1 (7.1)	
Active smoking at sampling (%)	4 (28.6)	6 (42.9)
Medication at sampling		
Aminosalicylates (%)	1 (7.1)	
Azathioprine (%)	5 (35.7)	
Corticosteroids (%)	0	
Infliximab/Adalimumab + Azahtioprine (%)	3 (21.4)	
Infliximab/Adalimumab + Corticosteroids + others (%)	1 (7.1)	
Azathioprine + others (%)	2 (14.3)	
Aminosalicylates + others (%)	1 (7.1)	
Infliximab/Adalimumab + others (%)	1 (7.1)	

CD: Crohn's disease.

Table 7 Description of the characteristics of UC patients and their healthy relatives.

Baseline clinical characteristics	UC (N=14)	Healthy relatives of UC (N=14)
Male/Female (%)	5/9 (35.7/64.3)	8/6 (57.1/42.9)
Mean age (SD) at samples collection	42 (11.7)	35.2 (15.8)
Median BMI (IQR)	24.3 (20.4-27.8)	24 (20.4-25.7)
Mean duration of disease (SD) at sampling	7.1 (6.1)	

Baseline clinical characteristics	UC (N=14)	Healthy relatives of UC (N=14)
Disease behaviour at sampling		
E1 proctitis (%)	3 (21.4)	
E2 left sided colitis (%)	3 (21.4)	
E3 pancolitis (%)	8 (57.1)	
Active smoking at sampling (%)	2 (14.3)	5 (35.7)
Medication at sampling		
Aminosalicylates (%)	10 (71.4)	
Azathioprine (%)	1 (7.1)	
Aminosalicylates + Corticosteroids (%)	1 (7.1)	
Aminosalicylates + Azathioprine (%)	1 (7.1)	
Cellcept (%)	1 (7.1)	

UC: Ulcerative colitis.

4.2.2 Description of the dataset

The processing of the cDNA obtained from the 111 fecal samples of the IBD cohort using MetaTrans yield a total of 565Gbp (mean=5.17, and s.d.=0.97) and recovered 2.84 billion of pair-end reads (mean=25.6 million, and s.d.=4.8 million, Figure A 1) of 101 bp in read-length. The processing runtime lasted a total of 656hours, roughly 27 days (10h/sample on average when running samples in parallel using multithreading; 2h/sample when running samples using one thread concurrently). We obtained an average of a 12% of rRNA, and of a 35.2% (s.d.= 8%) of reads that could be mapped to the MetaHIT-14 gene catalog (Figure A 2).

4.2.3 Dataset analysis

4.2.3.1 Gene count analysis

The analysis of the human gut microbial composition was initially reported by first studies by means of the number of unique microbial gene counts that were associated to gut bacterial richness. Authors like Le Chatelier et al. (Le Chatelier *et al.*, 2013) were able to associate obesity, a low-grade inflammation condition, with low microbial gene counts combined with a low microbial diversity. We thus believe that a first insight into this level of analysis can be useful as a first starting point to have a primary overview before going into downstream analyses.

We first assessed the stability of the fecal samples during their two collection time points (three months apart for healthy subjects, one year apart for patients who remained in remission or from baseline until they underwent a relapse). For this purpose, we analyzed the raw functional mapped gene counts, i.e. without functional annotation, and identified differences between groups by performing multiple comparison tests, paired (Mann-Whitney) and independent (Wilcoxon) tests, as appropriate (Figure 4.8, Figure 4.9, Figure 4.10).

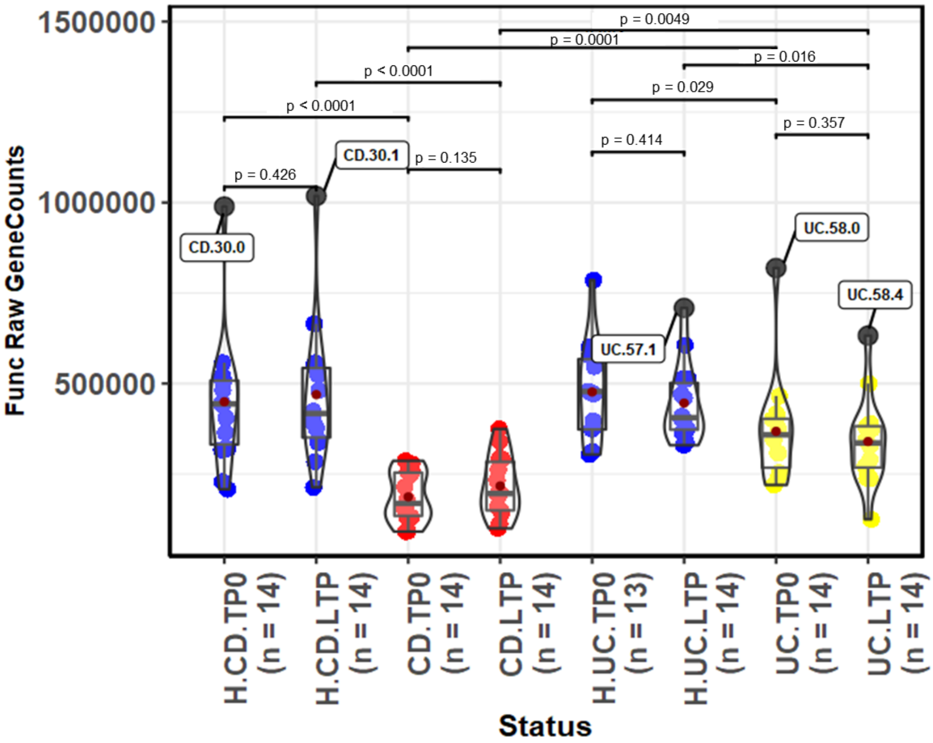


Figure 4.8 Between and within comparison of raw functional genes groups
Mann-Whitney tests for paired data between timepoints. Wilcoxon tests for the rest of unpaired comparisons. H: healthy relatives. CD: Crohn’s Disease. TP0: baseline. LTP: last timepoint. UC: Ulcerative Colitis.

Based on the number of genes (i.e. based on gene count). All groups (Healthy, CD and UC) did not present significant differences over time except for the group of CD in remission state at baseline (CD.REM.F.REM.TP0; F.REM stands for patients at baseline that remain in remission in the future, in last timepoint) and that remained in remission after one year (CD.REM.LTP); this group showed a moderate significance instability (p=0.0225, Figure 4.9; p=0.0759 trend, Figure 4.10). Nevertheless, CD and UC patients presented significant lower gene counts than their healthy relatives (Figure 4.8, Figure 4.9); this characteristic being more pronounced in CD than in UC. Gene counts of CD were also significantly lower than those of UC patients in both functional and taxonomical analysis only at basal timepoint, and interestingly higher in those patients that remain in remission state in both timepoints.

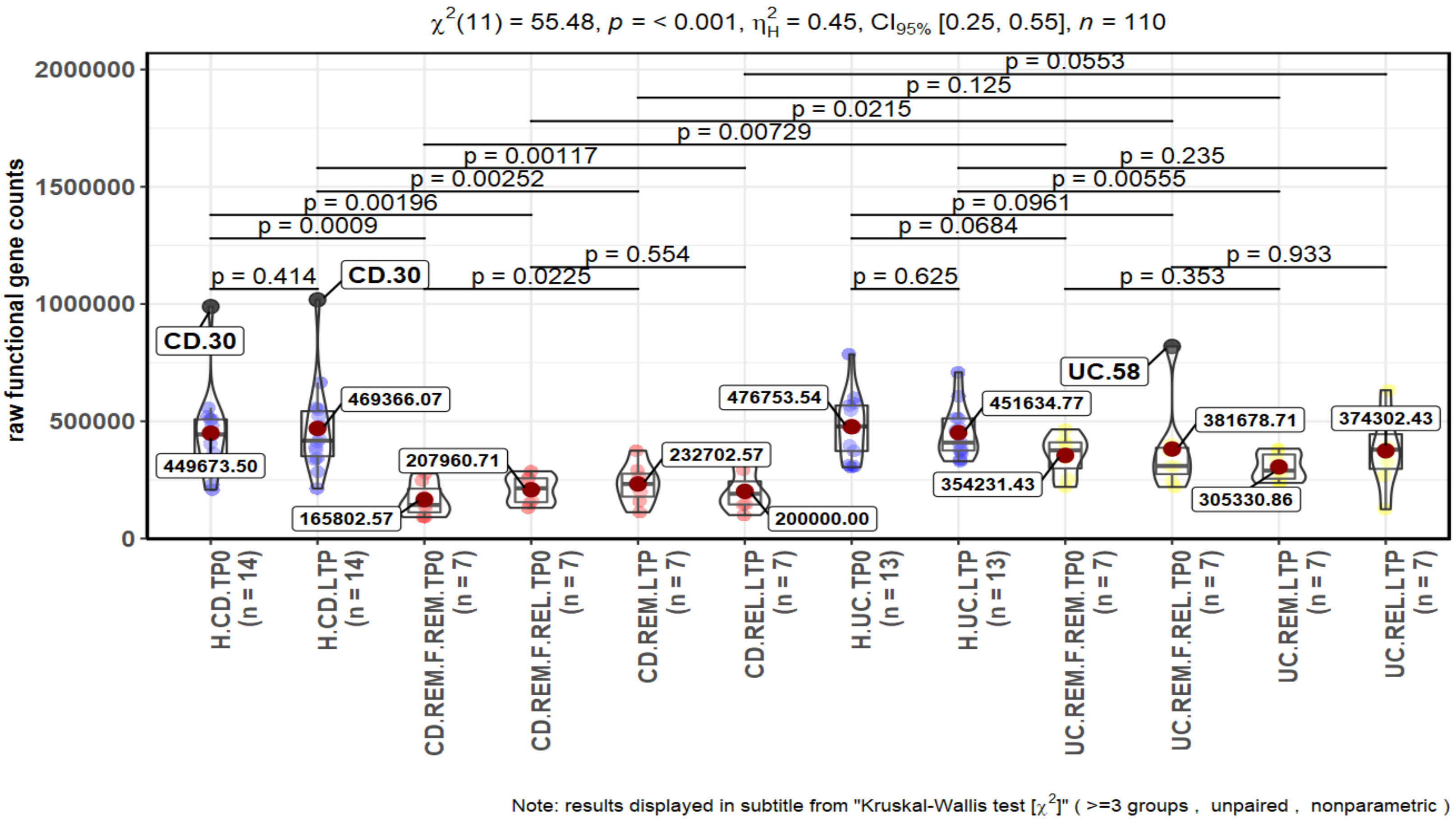


Figure 4.9 Between and within comparison of raw functional genes groups including REM/REL states.
Mann-Whitney tests for paired data between timepoints. Wilcoxon tests for the rest of unpaired comparisons.

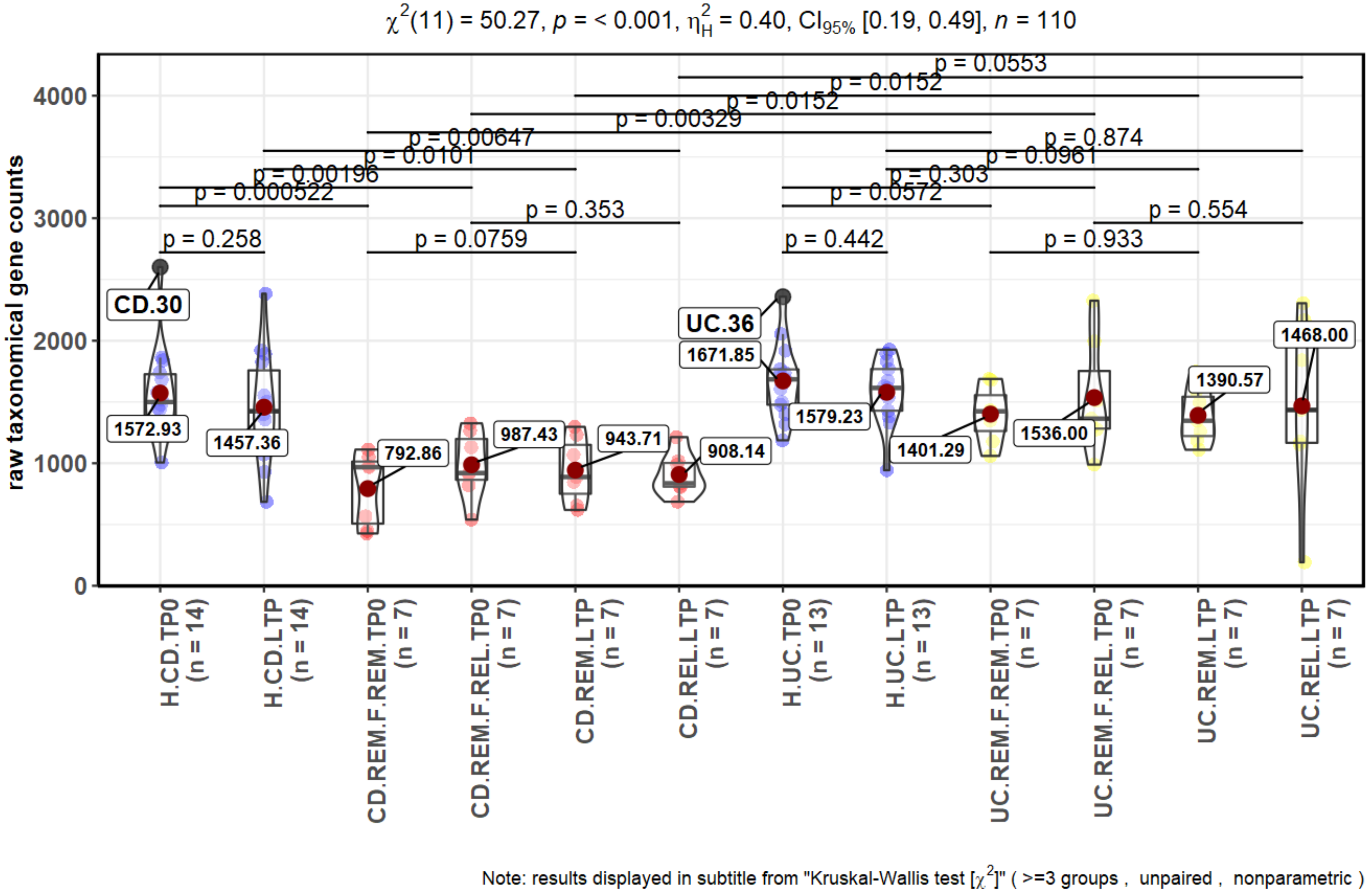


Figure 4.10 Between and within comparison of raw taxonomical genes groups including REM/REL states.
Mann-Whitney tests for paired data between timepoints. Wilcoxon tests for the rest of unpaired comparisons.

4.2.3.2 Functional and taxonomical annotation analyses

The following section makes use of what we called filters to find differences among groups. A filter refers to the comparison of two groups of samples specifically selected from functional or taxonomical profiling tables for testing a certain comparison. Therefore, filters allow selecting samples of interest in different scenarios of analysis. Each filter is described in detail in Figure 4.23 (at the end of this subsection) and is used as reference for the rest of the analyses. Please note that filter numbering is only referenced in case the reader needs to view in detail a filter, the comparison of interest is already mentioned within the paragraph.

Microbiome stability

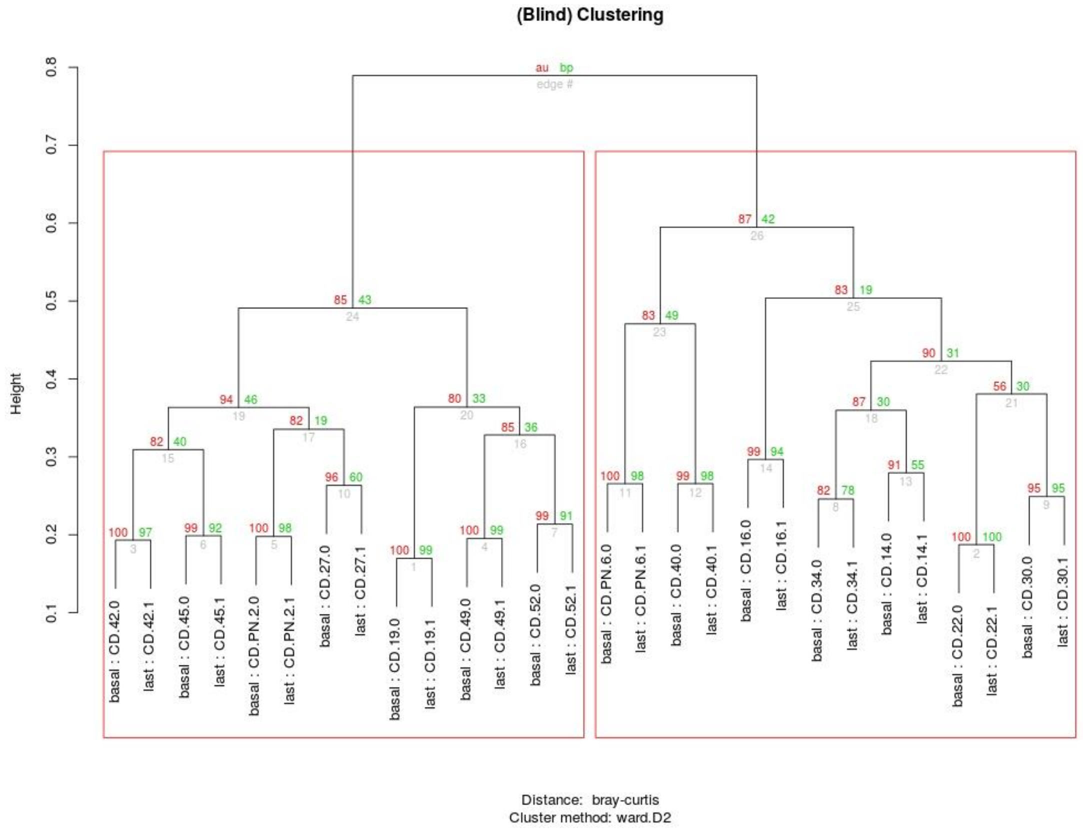
The stability of the microbiome along the two timepoints was analyzed by comparing healthy relatives of CD and UC patients (n(HR.CD)=14, n(HR.UC)=14) during basal and last timepoint (filter F1 and filter F3 respectively, Figure 4.23). The groups didn't show significant differences at any of the annotation levels of the functional and taxonomic databases. Same behavior was observed in the gene counts analysis (see previous section) which supports the hypothesis of a stable microbiome over time.

We further investigated differences in stability (F2) between groups of healthy relatives of CD (n=12) and healthy relatives of UC (n=13), but no statistical differences could be found among them. Given that relatives of patients did not differ, we could combine both healthy groups to increase power (group size; β-TypeII) when performing new filters.

During the analysis of stability in healthy CD subjects (F1), we found that most significant dendrograms did classify perfectly individual clusters of paired samples between time points at Genus rank and at Orthids EggNOGv4.5 functional annotation level, as shown in Figure 4.11.

At genus rank we obtained the highest overall p-values in the lower edges of the tree (AU p-values >0.95, at alpha=95%), which highlights a high clustering of the two time-point samples of each individual, and therefore suggests a greater inter-individual variability than intra-individual variability. A similar finding was obtained at the Orthids EggNOGv4.5 level, which also showed same classification of nodes (TruePositiveRate(TPR)=100%) at same significant level (p-values>0.95).

A)



B)

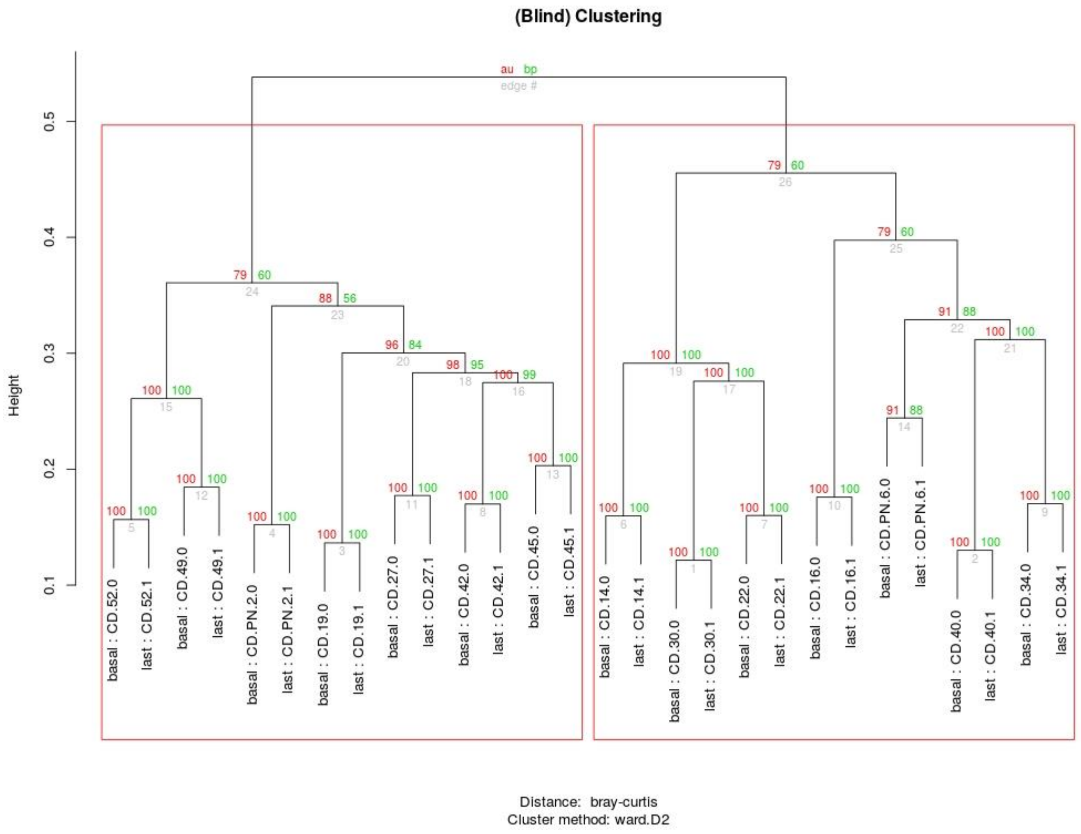


Figure 4.11 Dendrograms of healthy CD (H.CD) subjects at low taxa and functional annotation levels.
Comparison of hierarchical classifications of healthy CD subjects in both timepoints at low annotation levels (Genus rank in the taxonomical analysis(A) and EggNOGv4.5 orthologous genes(B)) with clusters at highest AU (approximately unbiased) p-values > 0.95 and significance alpha level 0.05.

As can be clearly appreciated in the dendrogram figures (Figure 4.11 A and B), the healthy CD group resulted in two differentiated groups based on significant differences of their microbiome at functional and taxonomical levels. A clinical metadata analysis confirmed these differences among ages (t-test, p=0.02; G1: mean=58 and s.d.=11.33, G2: mean=39 and s.d. =14.46), weight (t-test, p=0.02; G1: mean=65.7 and s.d. =10.44, G2: mean=85 and s.d. =15.6) and BMI (t-test, p=0.046; G1: mean=23.5 and s.d. =2.55, G2: mean=28.8 and s.d. =5.8). Hence, group1 conformed by samples CD.PN.6, CD.14, CD.16, CD.22, CD.30, CD.34, CD.40 can be described by having lower weight or BMI and higher age with respect to group2 identified by CD.42, CD.45, CD.PN.2, CD.27, CD.19, CD.49, CD.52.

The latter group was characterized by a significant enrichment at the family rank in *Bacteroidaceae* (21%, p-value = 0.0009, false discovery rate(FDR) of 0.004, Wald test*), Lachnospiraceae* (7%; p-value = 0.0002, FDR = 0.001, Wald test)*, Porphyromonadaceae* (2%; p-value = 0.008, FDR = 0.025, Wald test) and decrease in *Methanobacteriaceae* (2%; p-value < 0.001, FDR < 0.001), using abundance tables normalized by DESeq2 (via the median ratio normalization) (Love, Huber and Anders, 2014).

Performing a multivariate analysis of relative abundances at all annotation levels using PERMANOVA and Bray-Curtis distances, we did not observe significant results. However, an ordination analysis of genes using a Principial Component analysis (PCA), was able to separate both groups clearly at Phylum (54% variance in principal component 1), and Family ranks (21% variance in PC1) in the taxonomical analysis, and at orthologous ids annotation level in the functional analysis, “orthids” (29% variance in PC1) and “KEGG.orthids” (60% variance).

Conversely, in the analysis of UC healthy subjects we did not find a strong pattern of samples classification neither at a functional or taxonomic level, but we found same individual classification of samples between timepoints at low levels like species (TPR=93%) or orthologous EggNOGv4.5 (TPR=54%) annotation levels (AU p-values > 0.95). This suggests that there are no differentiated groups of UC healthy subjects, and samples are fully independent one of each other.

This initial analysis suggests that, in general, to perceive differences of clustering between samples we should focus more in lower levels (i.e. orthids, orthidsEggNOG4.5, KEGG.orthids,KEGG.funcat.L4, species, genus or family) whereas to have a broader picture of taxa/func composition we should use higher levels of annotation (i.e. functional categories, KEGG.L1, phylum, class).

Characterization of the active HEALTHY microbiome

To characterize the active microbiome of healthy subjects we selected the entire group of healthy subjects (n=55) independently of whether they were relatives of CD or UC (filter F4, healthy subjects). Count tables were sum-normalized to account for library depth and all features mapped without annotation were collapsed as "unknown". To obtain a better overview of patterns within the community, we collapsed all relative abundances lower than a 3% cutoff as "Other" category. The taxonomic analysis (appendix figures: Figure A 3-A, Figure A 3-C and Figure A 3-E) revealed a predominance of Bacteroidetes phylum (38%) over Firmicutes (21%), Actinobacteria (3%), Proteobacteria (1%) and Euryarchaeota (1%). Bacteroidetes and Firmicutes, which made up 60% of the total community, were therefore the most dominant bacterial division. *Unmapped* reads represented 27% (median, IQR=20%) of the data, with high variability between individuals ranging from 11% to 96%, highlighting the importance of unidentified bacterial organisms yet to discovered. At genus level (appendix figures: Figure A 3-B, Figure A 3-D and Figure A 3-F), we found a reduction of 14% (mean, s.d.=6%) in assigned taxa due to *unknown* assignments. The gut microbiome was primarily composed of Bacteroidetes (34%), gram-negative phylum, which included *Bacteroides* (24%), *Prevotella* (6%), *Parabacteroides* (3%), and *Paraprevotella* (1%), among others. These bacteria are well known to degrade food such as sugars (saccharolyticts) for the production of energy. The second dominant division, Firmicutes, consisted of *Blautia* (3%), *Ruminococcus* (2%), *Faecalibacterium* (2%), and *Roseburia* (1%), among others. *Collinsella* (1%) and *Bifidobacterium* (1%), two genera from the Actinobacteria phylum and *Methanobrevibacter* (1%), a genus from the Euryarchaeota phylum, accounted for lower proportions.

The functional analysis showed less variability and displayed a higher evenly distribution. At the functional categories level (EggNOGv3.0) (appendix figures: Figure A 4-A, Figure A 4-C Figure A 4-E), we observed a predominance of *unknown* functions (31%, which includes also uninformative categories "[R] General function prediction only" (7%) and "[S] Function unknown" (12%)), followed by functions related with *carbohydrate transport and metabolism* (12%), *[C] energy production* (8%), *[J] translation and ribosomal structure* (7%), and *[E] amino acid transport and metabolism* (6%). Another classification of functions could be done by means of the curated KEGG functional categories at level2 (appendix figures: Figure A 4-B, Figure A 4-D, Figure A 4-F) that contain a more spread-out classification of functions (46 detected functions compared to 25 in EggNOGv3). Again, the *unknown* category predominated (31%, including *Poorly characterized* (2%)), followed by functions with less than 0.3% of relative abundance collapsed into the *Other* (11%) category, *membrane transport* (8%), *translation* (7%), *energy metabolism* (5%) and *carbohydrate metabolism* (4%).

Nonetheless, a deeper analysis into the microbiome showed that most abundant annotation, functional or taxonomical, was not necessary the most prevalent in the group of healthy. The common microbiome core of functions and taxa describes more accurately those that are necessary for a bacterial survival, house-keeping genes, or for the gut ecosystem homeostasis. We, thus, used the relative frequencies to compute the most prevalent features at different relative abundances cutoffs, ranging from 0.1% to 10%. Then, we plotted a heatmap (appendix figures: Figure A 5) using the microbiome R package (Lahti and Shetty, 2012) and a minimum prevalence of 50%. This allowed the identification of core sets of taxa and functions that were described in detail in the appendix tables Table A 1-4, which are in line of previous studies of healthy population (Qin *et al.*, 2010; Li *et al.*, 2014; Lloyd-Price, Abu-Ali and Huttenhower, 2016; Rinninella *et al.*, 2019)

Healthy and patients

To assess statistical differences between healthy and patients, we initially joined samples from all healthy relatives of CD and UC and compared them against all patient samples of CD and UC (i.e. IBD) at remission (REM) and relapse (REL) states (filters F4-F5). Additionally, we also explored differences analyzing separately CD and UC samples, but comparing, inside each cohort, different combinations of sample groups (at basal/last timepoints and REM/REL status) (filters F6 to F21).

Comparison of the microbiome of all healthy relatives combined (n=55) with that of all IBD patients (n(UC+CD)=42) at baseline (i.e. under remission) and at relapse (REL) state (n(IBD.REL)=13) did not show significant differences at any functional or taxonomical levels.

Only a less conservative analysis using PERMANOVA with “rlog” transformation showed significance in both comparisons (F4, PERMANOVA, R2=9%, FDR=0.003; F5, PERMANOVA, R2=5%, FDR=0.003). An ordination analysis at high taxonomical and functional annotation levels showed a slight shift between both groups (Figure 4.12 A and B respectively, displaying variances >35% in their first principal component).

This lack of strong differences between healthy relatives and IBD patients might be explained by distinct microbiome compositions between UC and CD patients. Earlier studies have usually addressed comparisons between healthy controls and IBD patients as a single group. However, as it has been recently reported (Pascal *et al.*, 2017), UC microbiome composition resembles more to that of healthy than to CD microbiome. Such evidence could be the cause of an eventual interference in statistical signification. In order to obtain a stronger signal, we then decided to perform the rest of the analysis separating CD and UC.

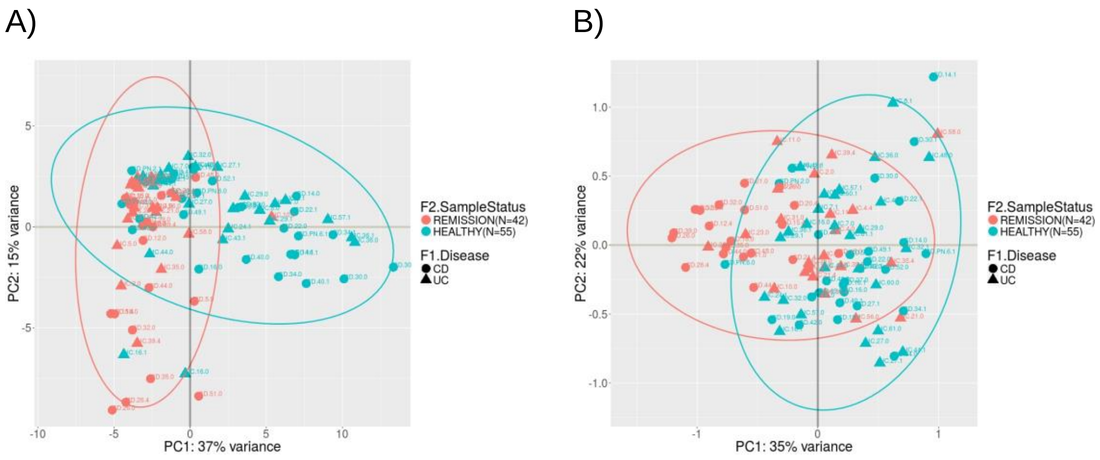


Figure 4.12 Principal component analysis (PCA) between healthy and IBD patients at remission state (filter F4).
Dimensionality reduction of taxonomical phylum annotation level (A) and functional categories annotation level (B) using the unsupervised database (EggNOGv3). The PCA was computed using Euclidean distances calculated from regularized log transformation counts.

All intragroup comparisons (i.e. baseline versus follow-up time point sample; healthy versus patients) of patients with healthy controls (filters F6-F13) did not show significant differences at any functional or taxonomical annotation level, except for the comparison between CD patients and their healthy relatives at baseline (filter F9; n(HR.CD.TP0)=14, n(CD.TP0)=14). This comparison presented significant results at functional (EggNOG[v3,v4.5] and KEGG[L1 to L4], with PERMANOVA, 20% ≤ R2 ≤ 30%, FDRs<0.05) and at gene count levels (filter F54, Figure 4.8, FDR= 0.0003, Wilcoxon test).

To perform a within-sample composition analysis (alpha-diversity) we calculated the Shannon index (diversity), Chao1 richness estimator, and the Pielou's evenness metric on the data annotated with the KEGG-functional database. The Shannon index showed a higher diversity (p=0.00097, Wilcoxon test) and a higher evenness (Pielous index, p=0.00042, Wilcoxon test) of the microbiome of patients under remission compared to healthy controls (Figure 4.13-B). These findings were similar using data annotated with the EggNOG database. However, using a taxonomic database, these differences between patients and healthy controls were not observed.

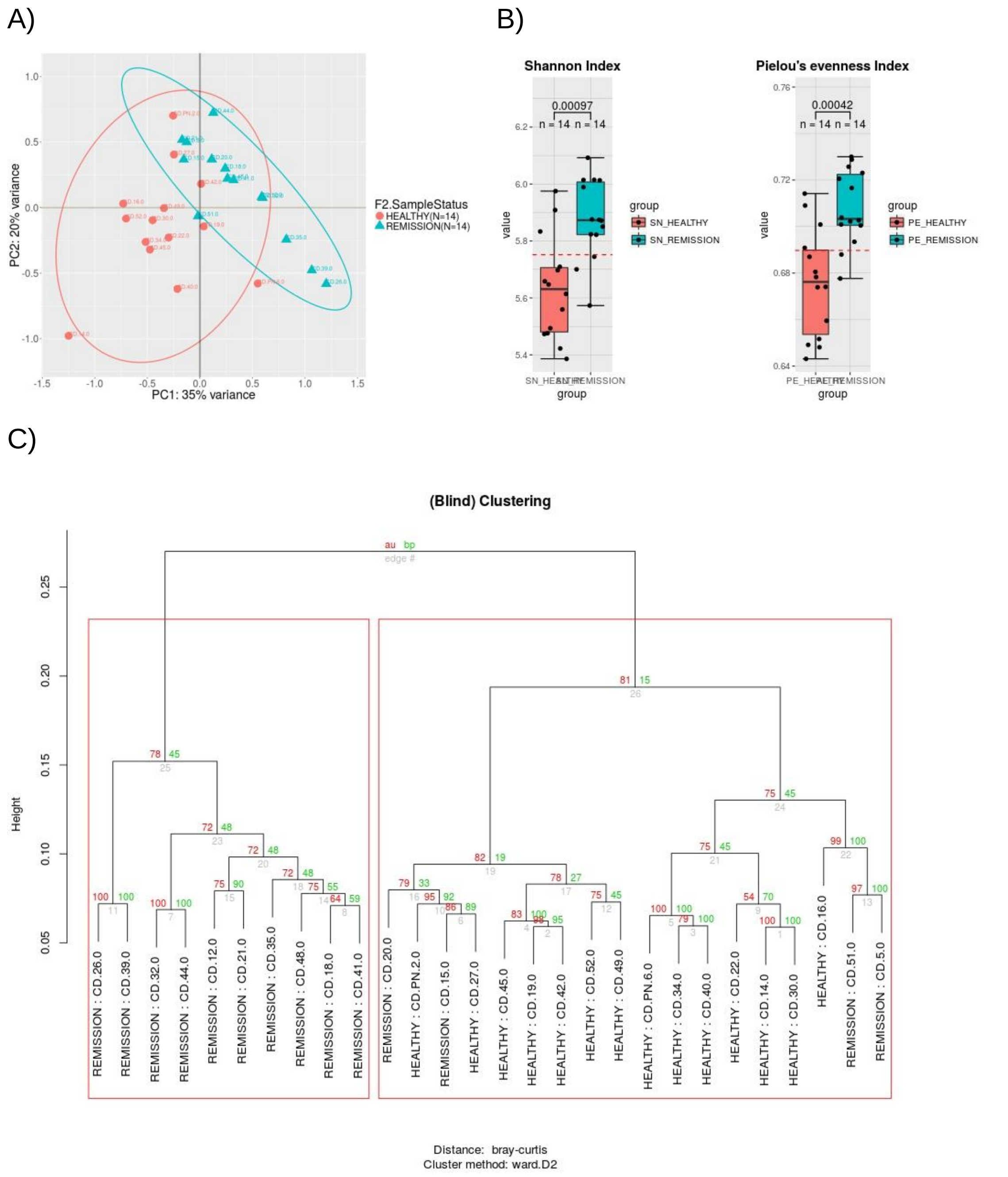


Figure 4.13 Functional beta and alpha diversity analysis between healthy relatives of CD and CD patients at basal timepoint.

CD patients under remission (REMISSION) are shown in blue. In red, healthy relatives of CD (HEALTHY). A) Principal component analysis (PCA) using "rlog" normalization as normalized counts to account for Euclidean distances. Ellipses groups samples by the 95% confidence of the population mean in the group. B) Boxplots of alpha-diversity measures calculated using sum-normalized counts (scale 0-1), Shannon and Pielou's indexes respectively. P-values are obtained from a Wilcoxon test between groups. Red doted line displays the mean of medians between groups. C) Hierarchical clustering (dendrogram) of samples computed using relative abundances and Bray-Curtis dissimilarities from orthologous KEGG genes. Green values on the edges indicates p-values computed using normal bootstrap resampling. Red values are p-values computed with a better approximation to unbiased p-values (Suzuki and Shimodaira, 2015). In grey, edge numbering. Red box displays the biggest partition in two groups found (k=2).

The clustering analysis (PCA) based on “rlog” distances delineated a consistent partition between both groups (Figure 4.13-A, PC1 variance = 35%), as well as using the non-parametric approach using Bray-curtis distances (NMDS). Additionally, using a hierarchical clustering with Bray-curtis distances we identified a clear classification into two distinct groups (Figure 4.13-C, see edges #25 and #26) with AU p-values between 75% and 100% in lower edges. Only four samples of CD where misclassified within the HR.CD (HEALTHY) group.

The differential expression analysis allowed us to identify 1,951 differentially expressed genes (DEG), with an FDR<0.05 and a log2 fold change (log2FC) average of 2 (s.d.=1.4) and -2.4 (s.d.=1.2) for up and down regulation respectively. Remission CD patients, compared to healthy controls, presented several upregulated functions at KEGG-L2 level that included *Metabolism of other aminoacids* (log2FC=0.5, FDR<0.0001), *Immune diseases* (log2FC=0.5, FDR=0.005), *Cardiovascular diseases* (log2FC=0.45, FDR=0.03) and *Nucleotide metabolism* (log2FC=0.2, FDR=0.02).

On the other hand, other KEGG-L2 functions were found downregulated in CD patients such as: *Digestive system* (log2FC= -1.2, FDR<0.0001), *Endocrine system* (log2FC=-0.5, FDR=0.004), and *Cell motiliy* (log2FC=-0.4, FDR=0.002). The most up/down regulated functions at KEGG-L2 matched with those components with higher and lower eigenvalues from the PCA analysis (Figure 4.13-A) which reinforced the idea that those functions were directly related with the differences between groups. We further investigated correlations at clinical level, but no correlations were found between clinical data and microbiome at KEGG-L2.

Additionally, this analysis also allowed us to identify significant (p-value<0.05) metabolic pathways over and under expressed (Table A 7, Figure 4.14). Top five most significative pathways were: Flagellar assembly (map02040, p=2.04e-19), Bacterial chemotaxis (map02030, p=8.2e-11), Peptidoglycan biosynthesis (map00550, p=1.12e-9), Methane metabolism (map00680, p=7e-9), and Carbon metabolism (p=3.4e-8).

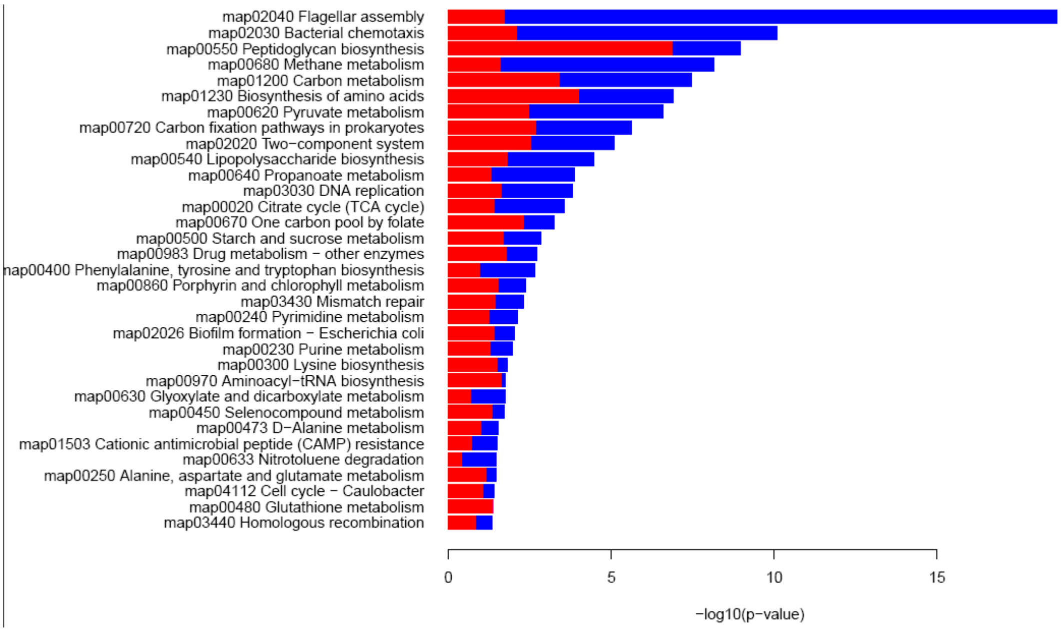


Figure 4.14 Barplot of most significant KEGG pathways, over and underrepresented, between healthy relatives of CD and CD patients (p<0.05)
Red and blue represent over and under abundance respectively.

Unlike CD, UC patients presented alterations of their active functions only at the level of EggNOG orthologous genes (orthidsEggNOG4.5) compared to their healthy relatives at baseline (filter F17; n(HR)=13, n(UC)=14; PERMANOVA, R^2=10%, FDR=0.032). UC patients also presented lower gene count than their healthy relatives (filter F56; p=0.029, Wilcoxon test, Figure 4.8). However, in alpha-diversity analysis, we did not find differences in diversity nor evenness, but only in richness, being lower in UC patients compared to healthy controls (Chao1 richness estimator, p=0.006, Wilcoxon test). A non-parametric ordination analysis (NMDS) using Bray-curtis dissimilarities did not show clear clusterization of groups, though we observed a shift between both groups in opposite directions in the first component NMDS1 (around 30% of patient and healthy samples were confounded). This observation suggests a weak dysbiosis in the UC microbiome.

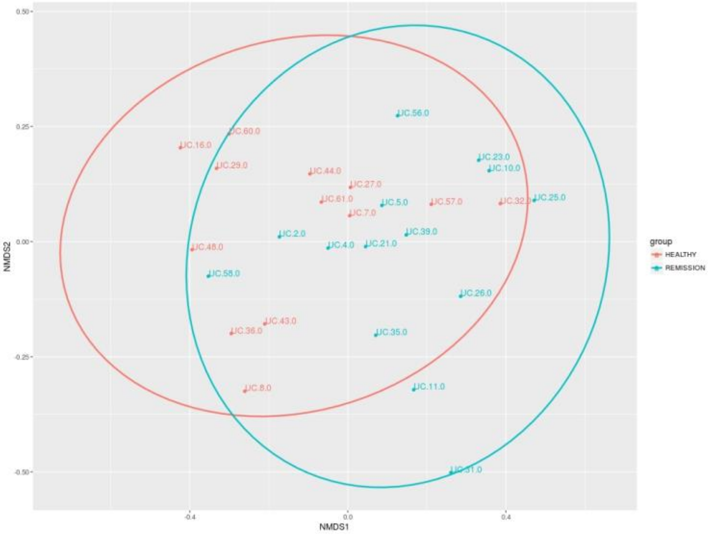


Figure 4.15 NMDS between healthy and patients of UC
Non-parametric multidimensional scaling (NMDS) graph using Bray-Curtis dissimilarities and orthologous EggNOG4.5 annotations as input. Healthy group is displayed in red. UC patients in blue. Ellipses display the 95% confidences ellipses for the population based on standard deviation (group's spread centroid).

Comparing healthy relatives of UC with UC patients we could identify 2,430 and 2,649 differentially expressed EggNOG orthologous genes up and down regulated respectively (FDR<0.05), with a corresponding log2FC average of 2.9 (s.d.=1.3) and -2.4 (s.d=0.7). These DEG were further annotated to their corresponding 25 functional categories (Tatusov, Koonin and Lipman, 1997) for better functional comprehension (Figure 4.16, Table A 8). The five functional categories that accounted for highest differential expression were: Cell motility ("N"), Cell cycle control, cell division, chromosome partitioning ("D"), Translation, ribosomal structure and biogenesis ("J"), Amino acid transport and metabolism ("E") and Nucleotide transport and metabolism ("F").

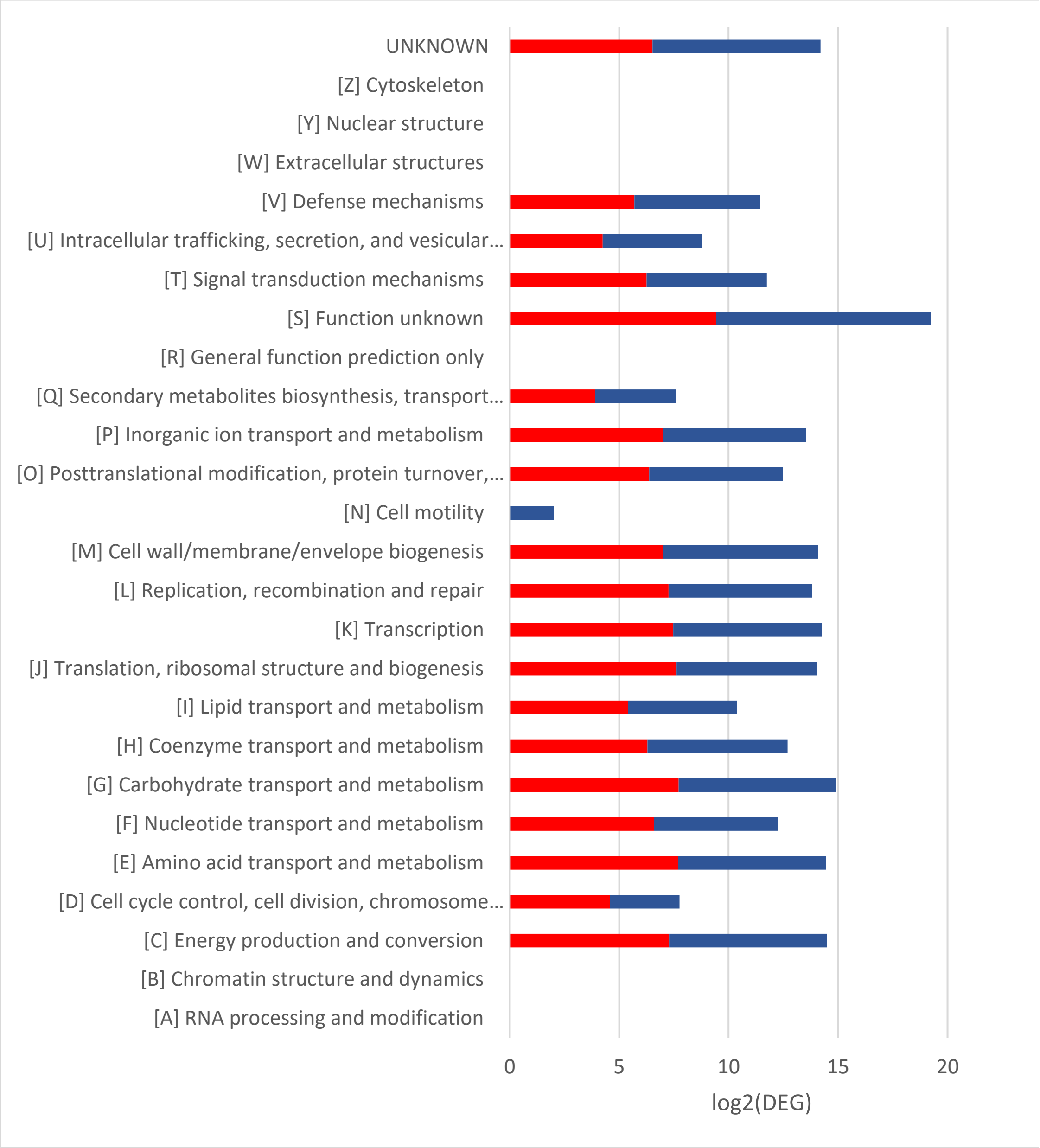


Figure 4.16 Barplot of differentially expressed EggNOG orthologous genes (FDR<0.05) of healthy relatives of UC compared to UC patients.
Red and blue represent over and under abundance respectively. Genes were annotated to their corresponding 25 functional categories according to EggNOG specifications.

When comparing CD with UC patients, we observed differences at species level either at baseline or at last timepoint (filters F27, F28 respectively; FDRs=0.081). Differences were also encountered at many functional (filter F27, at KEGG and EggNOG levels; FDRs < 0.05) and at gene count levels (filters F58, F9; Figure 4.8), with CD presenting a significant lower gene count than UC both at baseline and last time point.

Differences between samples (beta-diversity) employing 16S rRNA data between CD and UC patients were found higher at baseline (patients under remission; filter F27; n(CD)=14, n(UC)=14), compared to last timepoint were half of the patients underwent a relapse. The scatterplot visualization of the first two principal non-parametric multidimensional scaling (MDS) using Bray-curtis dissimilarity method (Figure 4.17-A) showed a concentric pattern fitting of UC samples, suggesting that microbial composition is not stronger enough to separate variances between groups. We also explored the distance-based tree to assess hierarchical clustering, resulting in an arrangement of samples that suggested a classification of CD and UC groups. An additional analysis of the alpha-diversity showed that diversity in UC patients was higher than CD (Shannon index, p=0.027, Wilcoxon test) (Figure 4.17-B)

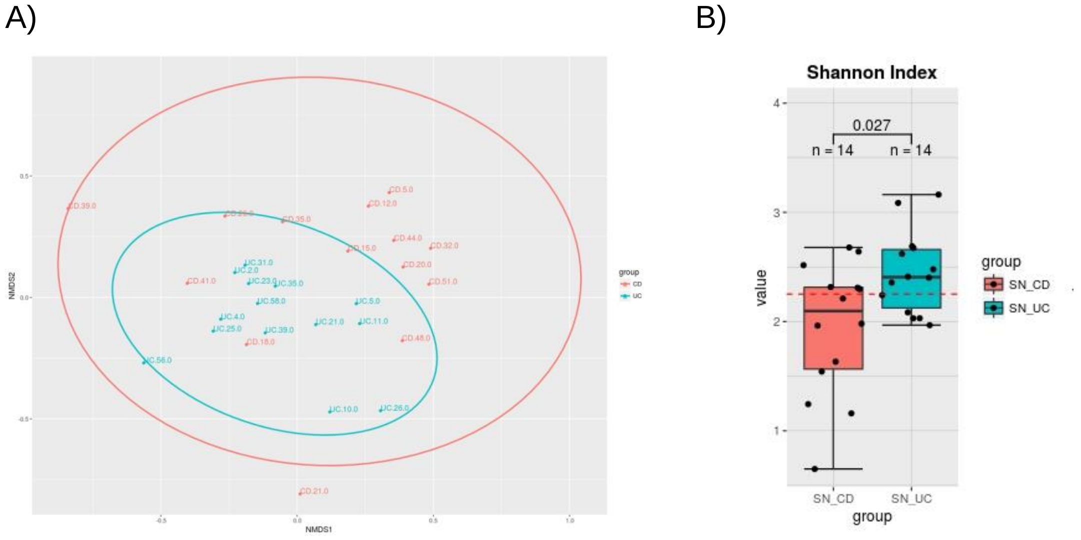


Figure 4.17 Taxonomical alpha and beta diversity plots between CD and UC patients at basal timepoint.
Comparison of 16S rRNA species between CD and UC at basal timepoint (filter F27; n(CD)=14, n(UC)=14). CD patients are shown in red. UC patients in blue. A) Non-parametric multidimensional scaling using Bray-curtis dissimilarities. Ellipses display the 95% confidences ellipses for the population based on standard deviation (group's spread centroid). B) Boxplot of Shanon diversity indices. Red dotted line shows the mean of medians in both groups. Significance obtained from Wilcoxon test between groups.

At baseline, the most upgregulated taxa in CD patients was *Fusobacterium* genus (log2FC=5.3, FDR=0.0003, Wald test) (Figure 4.18). This microbe has been previously observed in earlier studies as predominant in CD patients (Pascal *et al.*, 2017; Schirmer *et al.*, 2019). Additionally, we found a strong correlation of *Fusobacterium* with the gene CARD9.rs4077515 (rho=0.8, FDR<0.05). Other enriched species found were: *Ruminococcus gnavus*, identified as a prominent species in IBD (Hall *et al.*, 2017; Lloyd-Price *et al.*, 2019; Schirmer *et al.*, 2019), *Blautia producta*, associated with ileal CD (Walters, Xu and Knight, 2014), *Eubacterium dolichum*, abundant in non-high fat diet (Brown *et al.*, 2012; Liu, Qin and Wang, 2019). We also found a significant increase of *Bacteroides*, *Dorea* and *Prevotella* (significant only in the last timepoint, where remission and relapse are mixed; log2FC=5.2, FDR<0.0001, Wald test) genera. *Dorea* and *Bacteroides* have been observed in healthy subjects (Rinninella *et al.*, 2019), whereas *Prevotella* has recently been found increased in IBD (Lo Presti *et al.*, 2019).

Conversely, a reduction of genus *Eubacterium*, *Sarcina*, *Slackia*, *cc_115*, *Anaerostipes*, was identified as part of CD signature (Pascal *et al.*, 2017), *Bacillus*, and two uncultured *Paraprevotellaceae* and *Ruminococcaceae* families were found in CD patients compared to healthy controls.

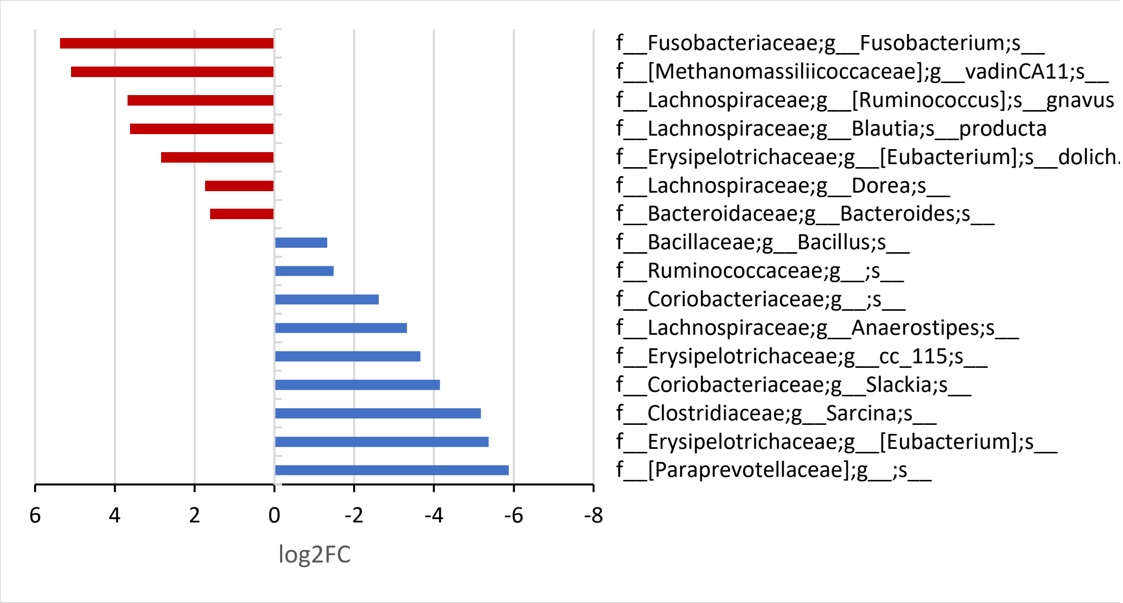


Figure 4.18 Differentially expressed taxonomy between UC and CD at basal timepoint

Red indicates over-expression, blue under-expression. Significance obtained at FDR<0.05 using the Wald test. Taxa annotated at family (f__), genus (g__) and species (s__) ranks.

In the last timepoint (filter F28) where half of the patients of CD and UC developed a relapse state, we encountered a depletion of the genus *Methanobrevibacter* (log2FC= -5.1, FDR=0.0001, Wald test) and the family *Christensenellaceae* (log2FC=-4.9, FDR=0.003, Wald test) that were identified as part of the CD signature (Pascal *et al.*, 2017). There were also two differentially low abundant species linked with depletion in IBD in earlier studies (Hall *et al.*, 2017; Lloyd-Price *et al.*, 2019): *Blautia obeum* and *Faecalibacterium prausnitzii* (known to be a butyrate producer (Kostic, Xavier and Gevers, 2014)) respectively. The most downregulated was an unclassified member of the family *Peptococcaceae* (log2FC= -7.1, FDR=0.00016, Wald test) and was found upregulated in UC patients in another IBD study (Van Der Giessen *et al.*, 2019). Furthermore, we also found a positive correlation of *Ruminococcus* and and uncultered lineage of the RF39 order with copies of *Akkermansia* (rho=1, FDR<0.05), and a negative correlation with *Blautia obeum* (rho=-1, FDR<0.05).

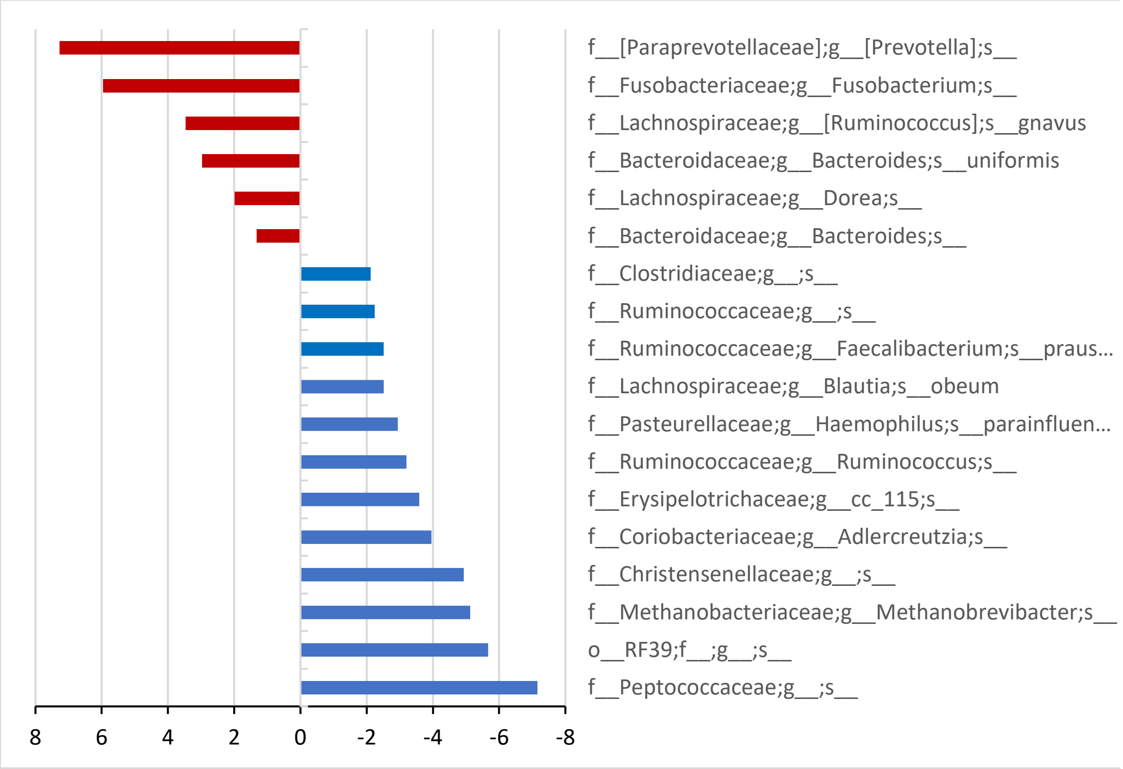


Figure 4.19 Differentially expressed taxonomy between UC and CD at last timepoint.
Red indicates over-expression, blue under-expression. Significance obtained at FDR<0.05 using the Wald test. Taxa annotated at family (f__), genus (g__) and species (s__) ranks.

At one of the lowest functional annotation KEGG levels, functional funcatKEGG.L4, we observed stronger differences between UC and CD patients at baseline. The non-parametric MDS ordination using Bray-Curtis dissimilarities showed significance in the non-parametric multivariate test (filter F27; PERMANOVA, R^2=16%, FDR=0.014). A similar approach using the “rlog” transformation and Euclidean distances supported also significance at this level, capturing a 21% of variance in the first component (Figure 4.20-A). A sample to sample heatmap of Euclidean distances using Ward.D2 as agglomeration method for linking and “rlog” counts transformation as input, showed a clusterization of CD and UC samples, supported by the hierarchical clustering analysis displayed on top with same distances (Figure 4.20-B).

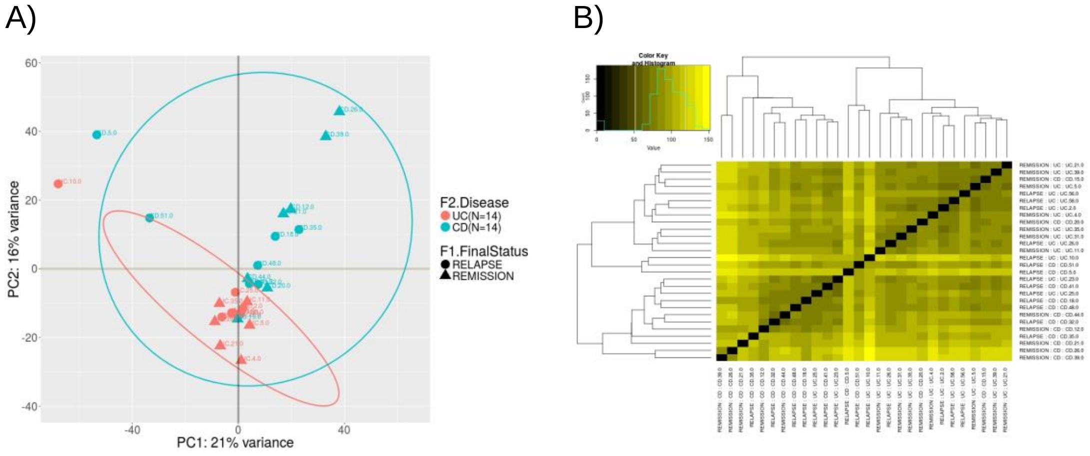


Figure 4.20 Beta diversity functional analysis between CD and UC patients at basal timepoint.
A) Principal component analysis (PCA) of regularized-log counts using Euclidean distances. UC patients in red. In blue CD patients. B) Heatmap of sample to sample distances from PCA plot. Yellow indicates higher distances between samples, black means closer. Most distant samples are clearly displayed in the cross, that separates groups, CD patients are mostly placed on the x-axis left, whereas UC are placed on the right. On top, dendrogram computed from same distances used in PCA and Heatmap, showing the hierarchical clustering.

Functional alpha-diversity indices showed lower diversity in UC patients compared to UC (Shannon diversity index, p=0.016, Wilcoxon test), the opposite in taxonomical analyses, where higher diversity is an indicator of a healthy microbiome. The Pielou's evenness index was also found significant (p=0.014, Wilcoxon test), suggesting that probably differences are not at richness level, but related with evenness.

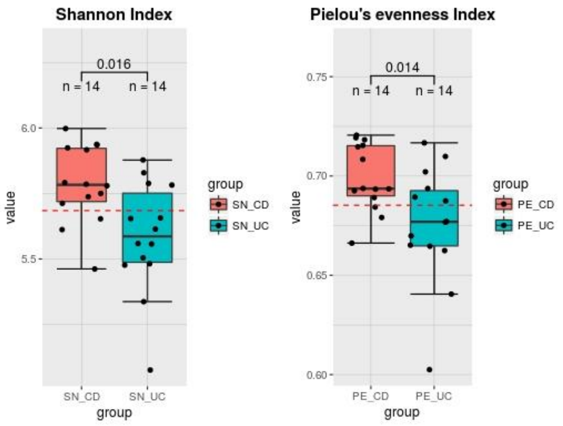


Figure 4.21 Functional alpha diversity indices of CD and UC patients at baseline.
CD is shown in red, UC in blue. Boxplots of Shannon diversity and PIelous' evenness indices obtained. Significance shown from a Wilcoxon test. Red line indicates the mean of medians between groups.

The comparison of CD and UC allowed us the identification of 690 and 438 up and down regulated genes (FDR<0.05) with a lower log fold change compared to previous comparisons (Up, log2FC mean=1.9 (s.d.=1.11); Down, log2FC mean=1.6 (s.d.=0.6)).In CD patients we found differentially expressed KEGG orthologous genes that allowed the detection of 34 pathways over or under abundant (Figure 4.22,Table A 9). The top five most significant were: Flagellar assembly (map02040), Bacterial chemotaxis (map02030), Peptidoglycan biosynthesis (map00550), Methane metabolism (map00680), and Carbon metabolism (map01200).

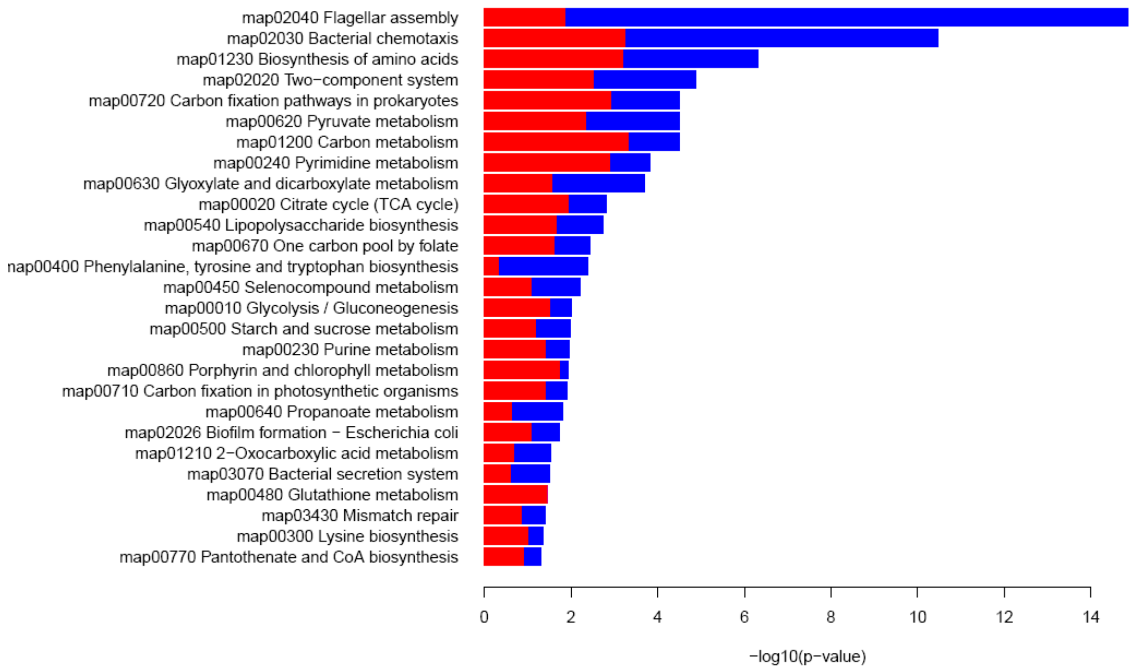


Figure 4.22 Barplot of most significant KEGG pathways, over and underrepresented, between UC and CD patients (p<0.05)
Red and blue represent over and under abundance respectively.

In other comparisons between CD and UC but including their corresponding healthy relatives in the comparison group, we observed differences between CD and UC, but only at functional levels and at low significance (filters F23, F25,F26; FDRs<0.07).

We did not observe significant alterations associated with severity of the disease, which may be attributed to the low number of relapse subjects (n=7 in each disease group). However, in the comparison between relapse and remission samples of UC patients we observed a weak significance (filter F34; PERMANOVA, R^2=8.3%, FDR=0.075) at functional level (KEGG-L4).

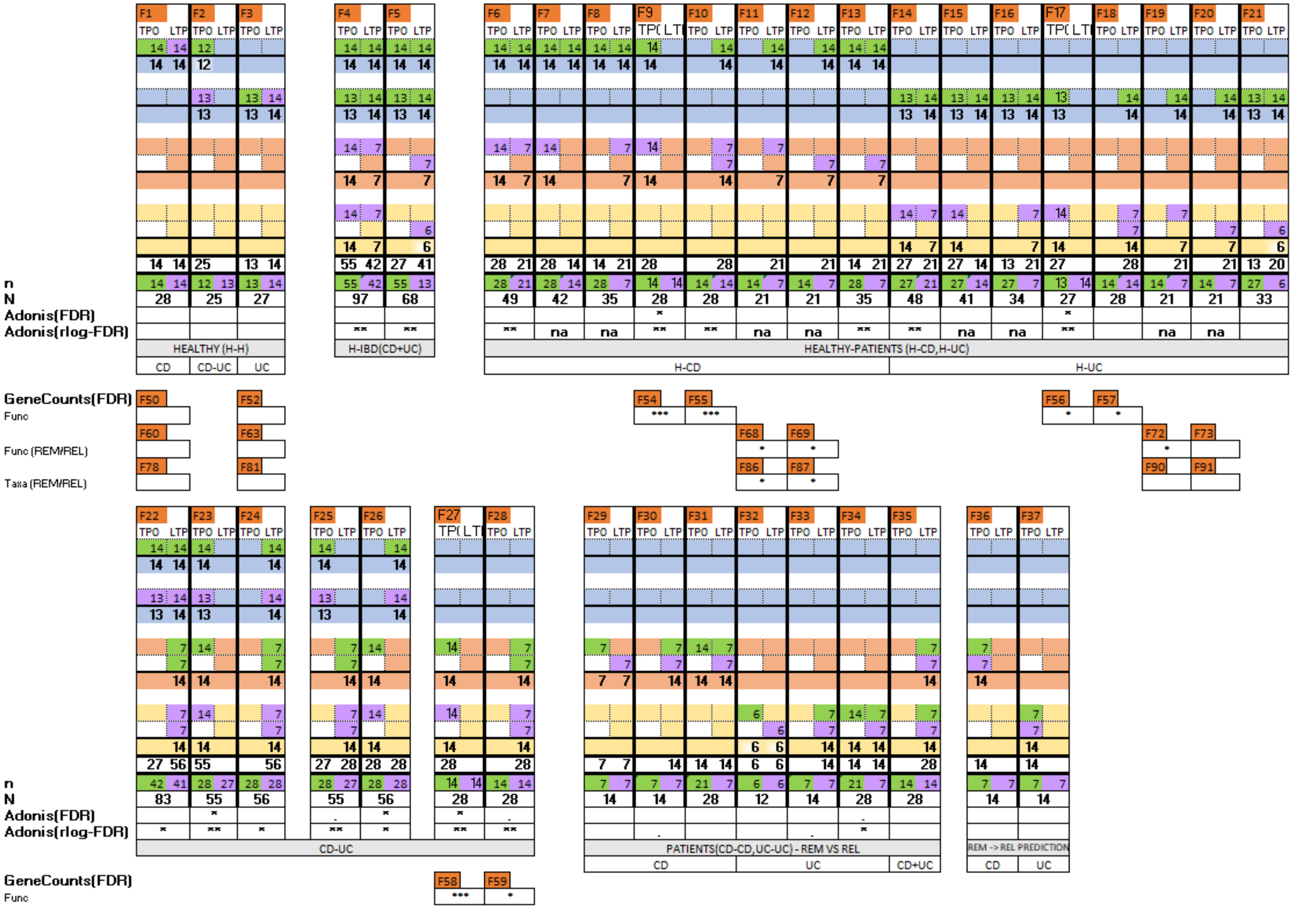

n
N
Adonis(FDR)
Adonis(rlog-FDR)
HEALTHY (H-H)
CD
CD-UC
UC
H-IBD(CD+UC)
HEALTHY-PATIENTS (H-CD,H-UC)
H-CD
H-UC
GeneCounts(FDR)
Func
Func (REM/REL)
Taxa (REM/REL)
CD-UC
PATIENTS(CD-CD,UC-UC) - REM VS REL
CD+UC
REM -> REL PREDICTION
TPO LTP

Figure 4.23 Reference of filters used for comparisons and Adonis test significance.
TP0 initial timepoint, LTP,last timepoint. First blue row, healthy CD subjects. Second blue row healthy UC subjects. Red row, CD patients; Yellow row, UC patients. Colored cells with numbers indicate the number of subjects taken in that set (green and purple cells are used to indicate groups used in the comparison). Statistical significance is shown for Adonis using sum-normalized counts and Bray-Curtis dissimilarities, and using “rlog” transformed Euclidean distances. GeneCounts comparisons from Figure 4.9 and Figure 4.10 are also shown. Statistical significance level defined as: * is <=0,05 (<=5%) ** is <=0,01 (<=1%) *** is <=0,001 (<=0.1%)

4.2.4 MetaTrans evaluation with HUMAnN2

Recently, in 2018, a new functional profiling tool developed by Eric Franzosa (Franzosa *et al.*, 2018) at the Huttenhower Lab (Biostatistics Department at the Harvard T.H. Chan School of Public Health) was published with the contribution of many top researchers in the metagenomics field such as J. Gregory Caporaso (Caporaso *et al.*, 2010), Rob Knight (Knight *et al.*, 2012), Curtis Huttenhower (Huttenhower *et al.*, 2012) and Nicola Segata (Segata *et al.*, 2011). This tool, namely HMP Unified Metabolic Analysis Network 2, is referred to as HUMAnN2. It is a pipeline for efficiently and accurately profiling the presence/absence and abundance of microbial pathways in a community from metagenomic or metatranscriptomic sequencing data and aims at describing the metabolic potential of a microbial community and its members.

This tool is unique as it provides as output not only a feature quantification (i.e. gene families, pathways, etc.) of the reads, but also provides a stratification for each feature which includes the taxonomical contribution; this is what they call "contributional diversity".

HUMAnN2 performs the analysis of DNA or RNA in two tiers. Briefly, the first tier maps reads against a taxonomical database with marker genes (~1M unique clade-specific marker genes, identified from ~13K prokaryote genomes among others) to classify them into identified well-known organisms. This task is performed using MetaPhlan2 (Huttenhower *et al.*, 2015). Once the microbial community is pinpointed, the reads are then mapped to a custom functional pangenome database named ChocoPhlan which represents all genomes used in MetaPhlan2 but annotated functionally using the curated comprehensive protein database UniRef50/90 (Suzek *et al.*, 2015). The remaining unmapped reads are later annotated using directly the UniRef50/90 protein database in a second tier.

The UniProt Reference Clusters databases (UniRef) provides clustered sets of sequences to obtain complete coverage of sequence space at several resolutions (100%, 90% and 50% identity) while hiding redundant sequences. UniRef90 and UniRef50 are built by clustering UniRef100 sequences at the 90% or 50% sequence identity levels respectively (Suzek *et al.*, 2015; Uniprot Consortium, 2019).

We performed a comparison of MetaTrans with HUMAnN2 to evaluate the differences. Eight randomly selected samples collected at basal timepoint from healthy subjects (n = 4) and Crohn’s disease (CD; n = 4) patients were used for this analysis.

The samples were processed in HUMAnN2 using multithreading (30 threads) in a dedicated server with 16 CPU cores and 128GiB of RAM. Each sample produced in average 30GiB (s.d.=14GiB) of data and consumed up to 16GiB in average (SD=8GiB) of RSS (resident set size) memory.

To fairly compare the functional annotation with MetaTrans (MT) output, we had to format the MH14 (MetaHIT-2014 or MH14) human gut gene catalog (Li *et al.*, 2014) to include it in the HUMAnN2 (HU) pipeline. We then conducted the comparison by processing the eight samples carrying out two types of analysis:

- Functional and taxonomical analysis (abbreviated “FT” analysis)
- Only functional analysis (abbreviated “F” analysis)

In both cases using these functional reference databases of genes and proteins (see Figure 4.24):

  - Gene database: MetaHit14
  - Protein database at 50% identity: UniRef50 (higher sensitivity)
  - Protein database at 90% identity: UniRef90 (higher specificity)

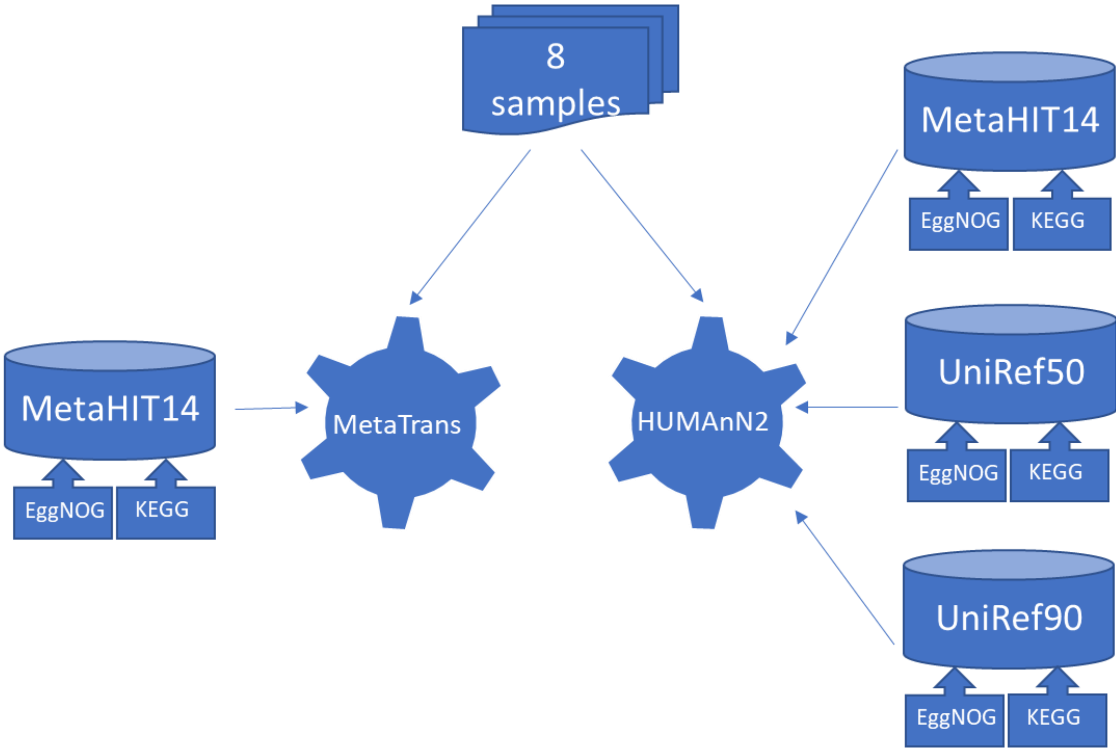


Figure 4.24 Comparison tools and database scheme.
Simple scheme summarizing the tools and databases used in the comparison between HUMAnN2 and MetaTrans. Protein database: UniRef50/90. Gene catalog database: MetaHIT14. Curated functional annotation database: KEGG Orthology database (KO) (Kanehisa and Goto, 2000; Kanehisa *et al.*, 2017, 2019). Non-curated and automated annotation database: EggNOG orthologous database (Huerta-Cepas *et al.*, 2016).

The percentage of mapping rates in HUMAnN2 outperformed MetaTrans by 16% ("MT.MH14" (mean=17.337e+06, and s.d.=7.493e+06) vs "HU.F.MH14" (mean= 23.173e+06, and s.d. =10.072e+06)) in the functional analysis, and roughly 20% ("HU.FT.MH14" (mean=25.498e+06, and s.d. =10.399e+06)) when performing an initial taxonomical assignment followed by a functional assignment (see Figure 4.25). In both cases the maximum achievement was obtained when using the MH14 gene catalog database.

We then conducted significance tests using the R package "ggstatsplot" (Patil and Powell, 2018). The Friedman test was used to detect significant differences between the approaches (p-value < 0.001). The Figure 4.25 illustrates the high differences in the percentage of mapping rates across configurations, which is appreciated by an effect size (W Kendall), close to one. The post-hoc pairwise tests revealed that all comparisons were highly significant (p-value < 0.001) with the exception of those marked as “ns” (non-significant). The significance shows a stratification between three differentiated groups, namely those performing an initial taxonomical assignment using more sensitive databases (“HU.FT.MH14”, “HU.FT.UniRef50”), those using only a functional assignment but using sensitive databases or performing taxonomical assignment but using high specific database (“HU.F.MH14”, “HU.F.UniRef50”, “HU.FT.UniRef90”), and finally those using higher specific databases or higher specific mapping settings (“HU.F.UniRef90”, “MT.MH14”).

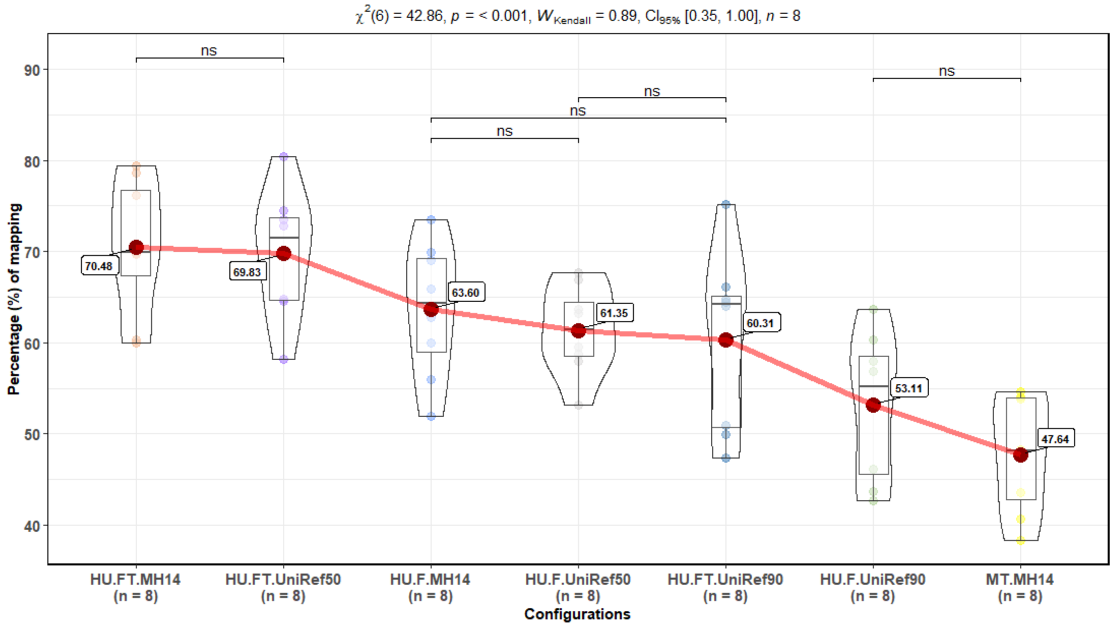


Figure 4.25 Mapping rates.
Boxplots and Violin plots of percent mapping rates. Significance: ns: non-significant, all pairwise not displayed have significance at level ***: p<=.001. Pairwise comparisons Durbin-Conover post-hoc test; Adjustment (p-value) Benjamini& Hochberg. Results displayed in subtitle from “Friedman rank sum test [χ^2]” (>=3 groups, paired, nonparametric). Red dots and red line show the mean of each group and the connection between them.

When we looked at the mean of genes recovered per sample, we observed a two-fold increase(x2.03) in HUMAnN2 at the functional analysis level using the MetaHIT14 gene catalog (“HU.F.MH14”, “HU.FT.MH14”) compared to MetaTrans (“MT.MH14”). Other settings using UniRef protein database achieved similar or a smaller number of genes than MetaTrans (Figure 4.26).

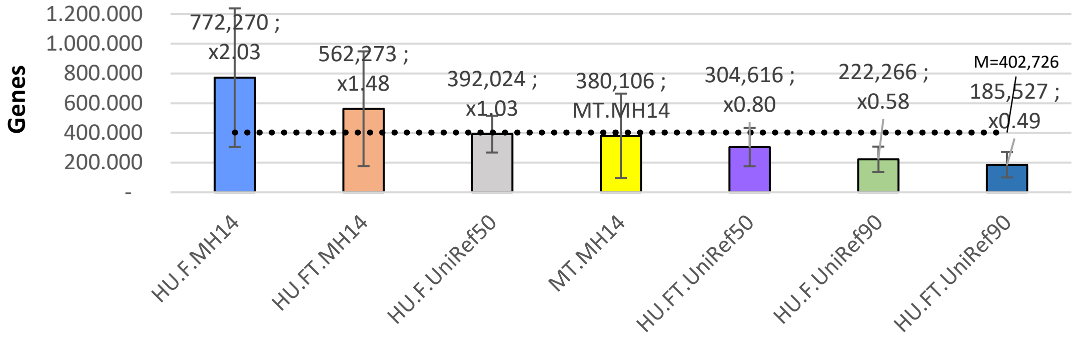


Figure 4.26 Mean of identified genes per sample.
Bars display means per sample. Whisker lines indicate standard deviation from the mean. Values next to mean values on top of bars indicate the fold change with respect to "MT.MH14" configuration. Dotted black line indicates the grand mean.

However, the number of unique genes recovered per configuration showed a decrease in fold-change difference (x0.69) between MetaTrans ("MT.MH14") and HUMAnN2 ("HU.F.MH14", "HU.FT.MH14"). This difference represents an increase of 30% in the number of unique genes recovered by HUMAnN2 when compared to MetaTrans. Interestingly, the number of unique genes recovered by HUMAnN2 performing the taxonomical analysis ("HU.FT.MH14") was less than without performing it ("HU.F.MH14") (Figure 4.27).

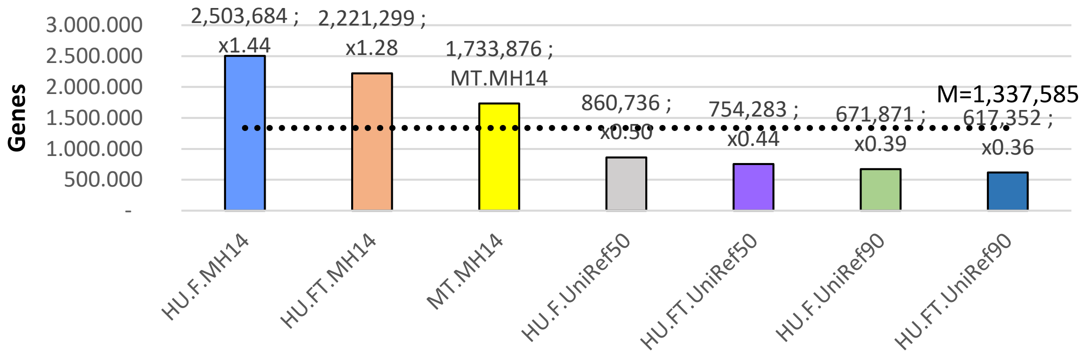


Figure 4.27 Number of unique genes recovered in each configuration.
Bars display the number of total unique genes per configuration. Values next to total values on top of bars indicate the fold change with respect to "MT.MH14" configuration. Dotted black line indicates the mean.

One of the main issues encountered when performing the functional analysis was the percentage of functionally annotated genes. As Figure 4.28-A and Figure 4.28-B display, the annotation percentage ranged from 73-10% in KEGG

(Kanehisa and Goto, 2000; Kanehisa *et al.*, 2017, 2019)(Figure 4.28-A) and from 78-31% in EggNOG (Huerta-Cepas *et al.*, 2016) (Figure 4.28-B). The difference in percentage of functionally known genes/proteins database is mainly driven by two differentiated groups in both figures, those that use the MetaHIT14 gene catalog and those using UniRef. We assessed statistically this difference with a Mann-Whitney U test to compare two independent non-parametric groups. The test indicated that the difference between the two groups was significant at 5% significance level (U=0, p-value = 0.05) in Figure 4.28-A, and almost significant (U=0, p-value = 0.057) in.Figure 4.28-B. The mean of percentage growth between the two groups when using the MetaHIT14 gene catalog was 48% in Figure 4.28-A and 33.6% in Figure 4.28-B.The difference observed in HUMAnN2 with respect to MetaTrans reflects and increase of the 13% and 24% when performing the initial taxonomical mapping in the case of the KEGG annotation database (Figure 4.28), and an increase of the 15% and 20% when performing the initial taxonomical mapping in the case of the EggNOG annotation database (Figure 4.28-B).

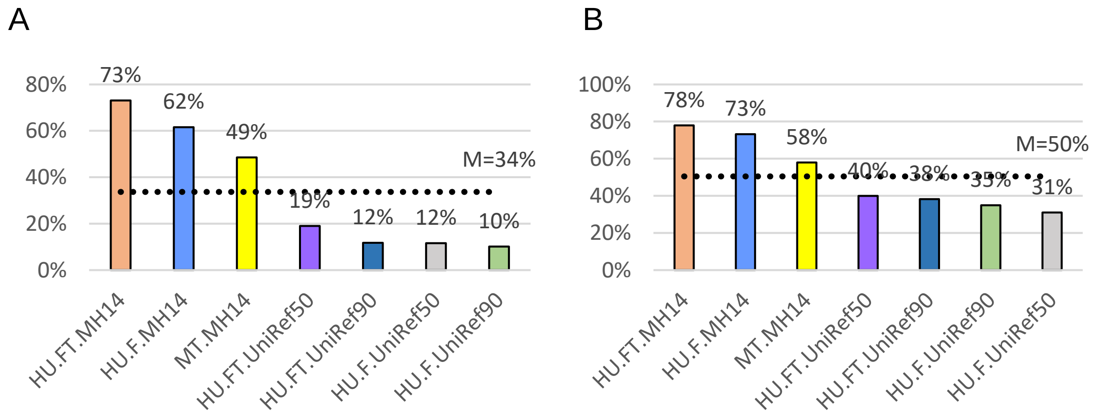


Figure 4.28 Known functional annotation.
Bars display the percentage of annotated functions per configuration. Dotted black line indicates the mean. **A** Annotation in the curated KEGG Orthology database (KO). **B** Annotation in the automated (non-curated) EggNOG orthologous database.

The analysis of the known annotated functions of the genes or proteins in the annotation databases highlights the lack of current annotation of almost half of entries in each database. The identification of genes or proteins is, therefore, not linked necessary to its known function. This fact reduces much more the pool of potential functions that are available to be identified in samples.

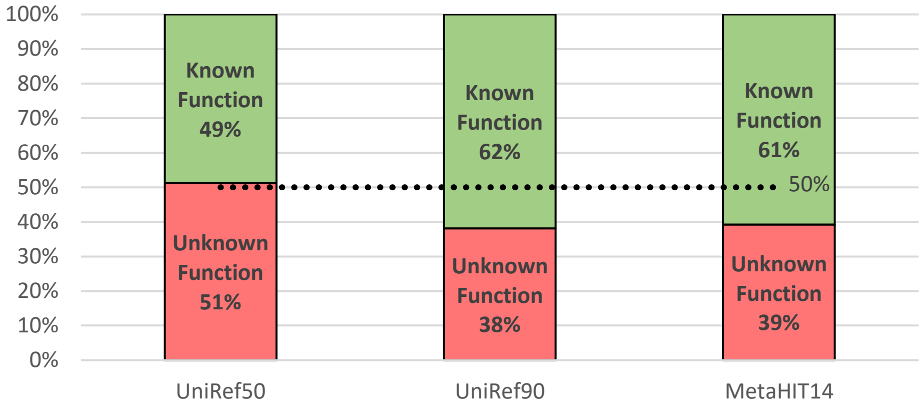


Figure 4.29 Known/Unknown functional annotation percentages in gene and protein databases.
Unknown functions are considered all annotations with the “unknown” label in the MetaHIT14 gene catalog, or with “uncharacterized protein” in the UniRef protein database.

The Figure 4.30, illustrates the runtime improvement achieved in HUMAnN2 compared to MetaTrans. The most comparable configuration to MetaTrans, according to the analysis type and mapping database used, is the “HU.F.MH14”. It obtained a decrease of x2.50 fold in terms of runtime, whereas when compared to “HU.FT.MH14”, which performed an initial taxonomical assignment, the runtime dropped to a fold change of almost four, x3.89.

To further analyze the difference between HUMAnN2 and MetaTrans, we assessed the comparison at different annotation levels depending on the functional reference database used. A non-parametric analysis (N=8) was conducted by means of Spearman's rank correlation coefficient (ρ) in each comparison. We considered only those configurations where only a functional analysis was performed (i.e. “HU.F.MH14”, “HU.F.UniRef90”, “HU.F.UniRef50”), and were compared against MetaTrans (“MT.MH14”). Matrices of ρ values obtained for each comparison were represented in heatmaps by mapping ρ values (1 to.-1) to colors (1-blue, highest positive correlation, 0-white, no correlation, -1-red, highest negative correlation), and p-values were calculated by each coefficient value. Cells were marked with a cross if p-values were not significant (p-value<0.05) (see Table 8).

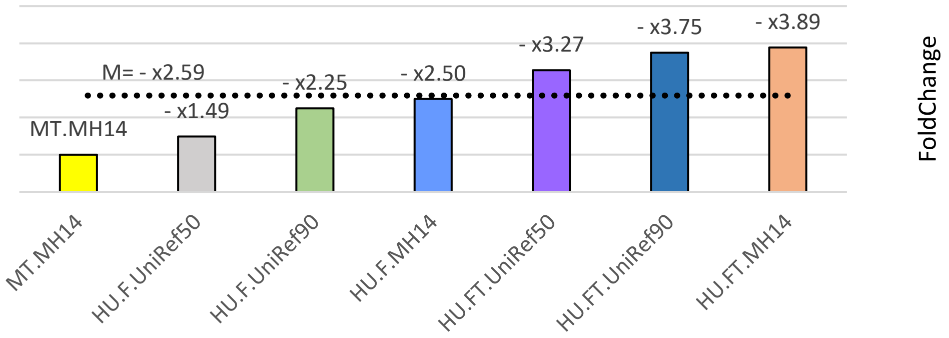


Figure 4.30 Runtime fold changes.
Bars display the runtime fold-change with respect to MetaTrans (“MT.MH14”). Dotted black line indicates the mean

The Table 8 depicts all possible annotation configurations (rows) by each functional analysis configuration (columns) compared to MetaTrans functional analysis. The major patterns of correlation were yield in the HUMAnN2 configuration using the same functional annotation database (“HU.F.MH14”) as MetaTrans, i.e. MetaHIT14 column, where diagonals (comparison of same sample) displayed always highest positive correlation values. In general, as the annotation level diminish correlation values decrease, this effect is caused by the fact that lower functional annotations (either in KEGG or EggNOG) became more specific and, therefore, differences were sharpened. When we compared between the three configurations, specificity was higher in the protein database UniRef90, where correlation values tended to be lower, and diagonals devise smoother correlation values. Overall comparisons of healthy to healthy samples show higher correlation values than comparisons of patients or healthy to patients.

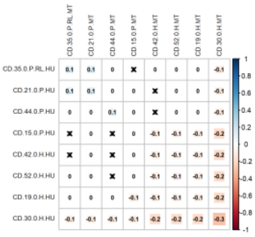
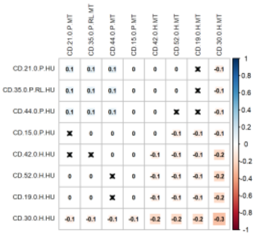
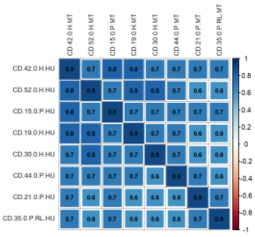
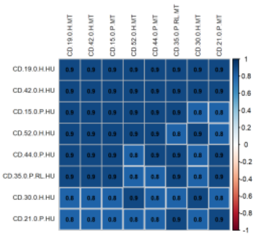
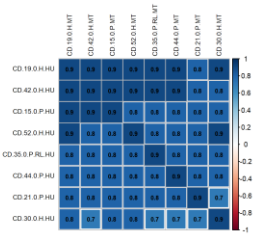
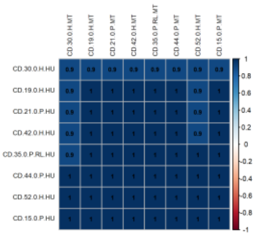
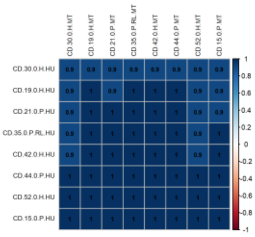
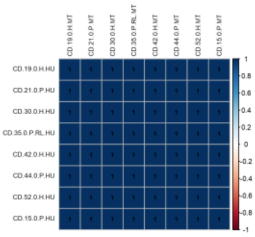
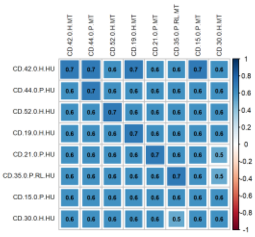
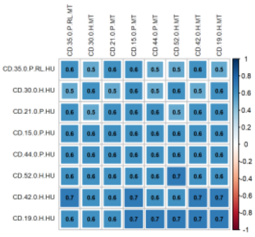
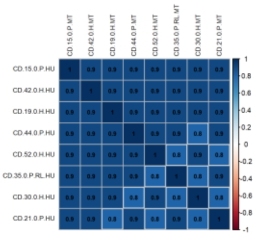
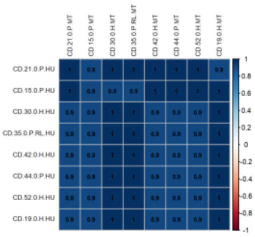
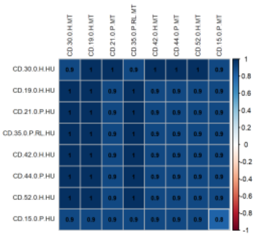
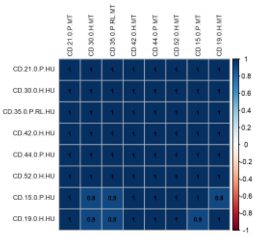
A particular case was observed in the EggNOG annotation database at orthologous genes id level (“EggNOG-OG.id”) when comparing between different functional protein databases (“UniRef50”, “UniRef90”). The low and negative correlation values (ranging between $0.1 \leq \rho \leq -0.3$) indicated a particularity when performing annotation of proteins to EggNOG non-supervised orthologous groups (NOGs), due to the difference version of EggNOG used in HUMAnN2 and MetaHIT14, whereas in the former they used EggNOGv4.5, in MetaHIT14 the version 3.0 was used. This difference implied a change in the NOGs ids format, and while functions were comparable at higher level (functional categories (Tatusov, Koonin and Lipman, 1997), “EggNOG-OG.funcat”) they were not at gene orthologous level.(“id” level). Thus, while one sample in HUMAnN2 has annotation values for one particular id, the same sample in MetaTrans has zero abundance, which explains the negative correlations observed. Therefore, in order to remove the bias in annotations, we removed non-common identifiers (“EggNOG-OG.shared.ids” row), which generated high positive correlations (between 0.8 and 0.9) as expected.

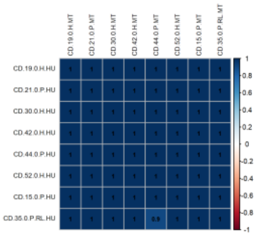
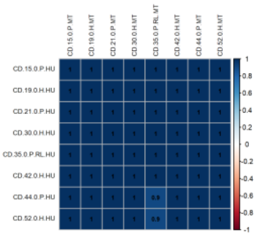
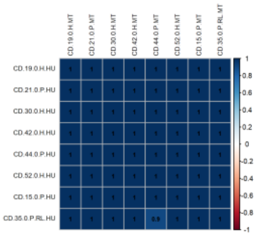
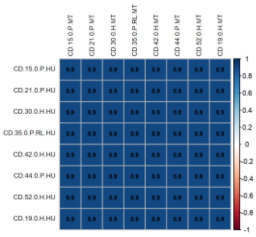
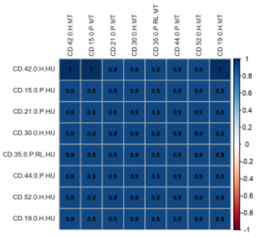
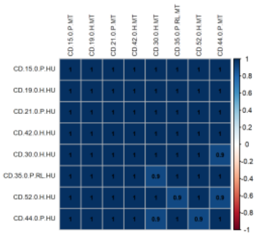
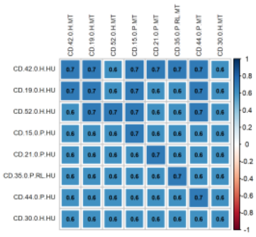
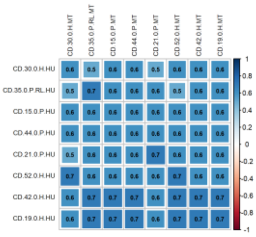
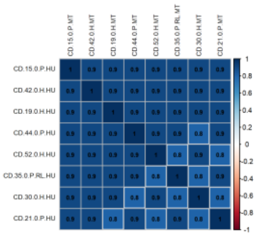
Therefore, we concluded that:

1. Functional outputs using MetaTrans could be considered comparable to HUMAnN at different levels : EggNOG-OG.ids, EggNOG-OG.funcat, KEGG-KO.L1 to KEGG-KO.L3, since the analyses showed a correlation >= 0.8.

2. At lower functional levels such as KEGG-KO.id and KEGG-KO.L4, correlations were between 0.6 and 0.7, therefore interpretations of the functional analysis should be taken with caution.

Table 8 Heatmaps of correlation matrices.

	HU.F.UniRef50	HU.F.UniRef90	HU.F.MH14
EggNOG-OG.id			
EggNOG-OG.shared.ids			NA
EggNOG-OG.funcat			
KEGG-KO.id			
KEGG-KO.L1			

	HU.F.UniRef50	HU.F.UniRef90	HU.F.MH14
KEGG-KO.L2			
KEGG-KO.L3			
KEGG-KO.L4			

Column and row headers of this table indicate, the HUMAnN2 configurations used to compare against MetaTrans and the functional annotation database level that is being compared, respectively. Within correlation matrices, rows are HUMAnN2 samples, columns samples analyzed in MetaTrans. Cells are marked with a cross if found not significant (p-value < 0.05). Correlation coefficient values (ρ) are displayed in a color scale (blue = 1, maximum correlation; red = -1, maximum negative correlation; white = 0, no correlation). Same scale is applied to cell color squares, higher correlation, bigger square size (either positive or negative correlation).

Interestingly, when comparing normalized abundances of functional categories (Tatusov, Koonin and Lipman, 1997) from HUMAnN2 using the MetaHIT14 gene catalog ("HU.F.MH14") and MetaTrans, displayed very similar proportions of functional assignment (Figure 4.31). This result further validates the use of MetaTrans at functional categories level.

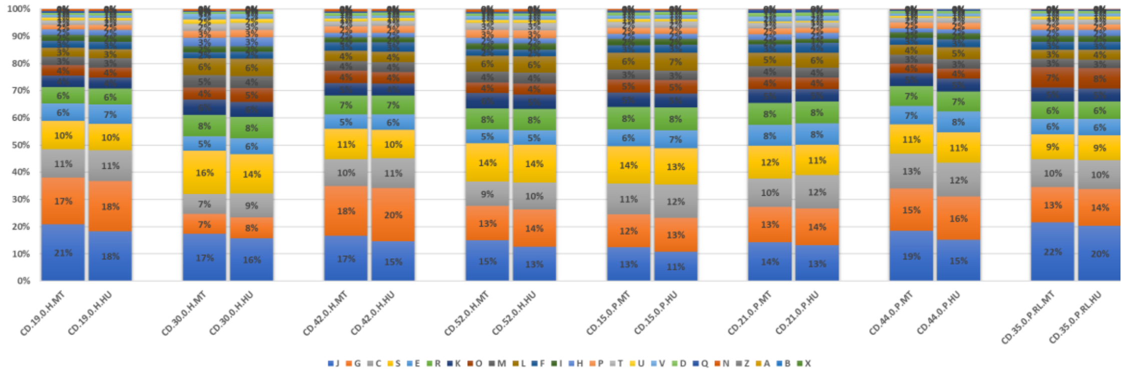


Figure 4.31 Percent stacked bar chart of EggNOG functional categories
Proportion of abundances at EggNOG functional categories annotation level from different samples analyzed in HUMAnN2 ("HU.F.MH14") and MetaTrans ("MT.MH14") using the MetaHIT14 gene catalog database.

Assessment based on previously described metrics between all configurations demonstrate a superior performance in those configurations mapping against MetaHIT14 gene catalog, followed by configurations performing a pre-taxonomical mapping using UniRef protein database and finally configurations performing functional assignment with UniRef. Additionally, assignments using UniRef database have slightly higher assignment and annotation with the UniRef-50, as expected due to the higher sensitivity of this database (Figure 4.32).

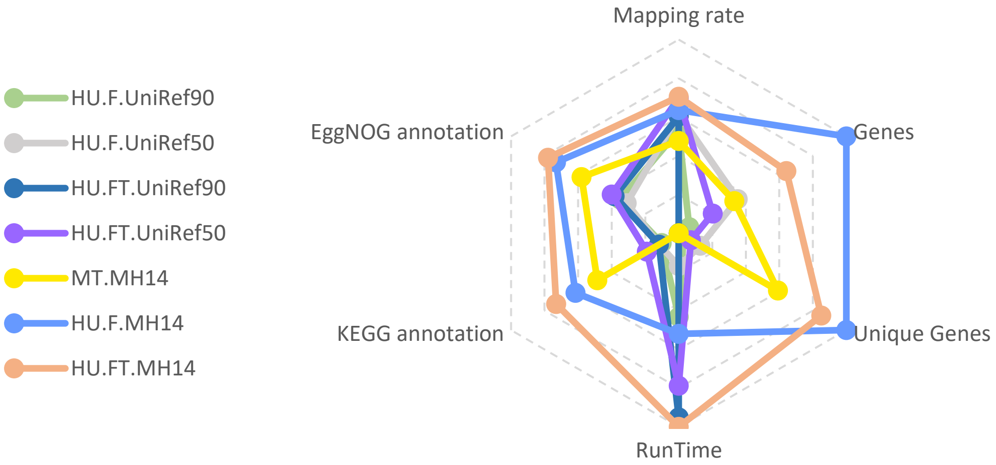


Figure 4.32 Radar chart summarizing configuration characteristics.
Summary radar chart reflecting, per configuration, the mean of percentage mapping rates, the mean of genes, number of unique genes, proportion of annotated functions in KEGG and EggNOG databases, and runtime fold-changes compared with MetaTrans as baseline. Metrics that do not represent proportions were scaled up to get relatives values by considering the minimum and maximum values as the lower and upper bound in the relative values. The further towards the edge of the spoke a point reaches, the higher mapping rates.

Chapter 5.

General discussion

5 General discussion

The objectives of this thesis were to address two main aspects. First, the development of a bioinformatics pipeline to analyze microbial cDNA sequences and, second, to apply this tool to characterize the active microbiome of healthy individuals and patients with IBD.

At the time we started this dissertation, we found very few published works handling microbial cDNA data obtained from total microbial RNA extracted from fecal samples. Most of the published tools addressed the cDNA obtained from Eukaryotic cells and therefore, were not appropriate to perform microbial taxonomic and functional analyses. Further, of the few tools we found, like HUMAnN (Abubucker *et al.*, 2012), one of the main issues we found was how to align roughly 25 million of paired-reads per sample in a reasonable time using free aligners (NCBI BLAST was very slow, and the other proposed options were paid-alternatives). Other metatranscriptomic analysis pipelines available were only available online, like MG-RAST (Meyer *et al.*, 2008) limiting our control over the samples and their analysis. Based on a few papers such as Gosalbes et al. (Gosalbes *et al.*, 2011) and Leimena et al (Leimena *et al.*, 2013), where they provided their analysis workflow in more or less detail, we designed and implemented a bioinformatics pipeline that overcame the aforementioned main issues, and focused on the human gut microbiome (Li *et al.*, 2014).

We used two human cohorts to validate (IBS cohort) and to apply (IBD cohort) our pipeline.

The results of the 16S rRNA analysis, which characterizes active bacteria, contrasted with those of 16S rDNA, thereby indicating that not all microorganisms identified at the DNA level play an active role in the gut community. Furthermore, active microbes such as Bifidobacteriaceae showed an increase in relative abundance as an effect of a flatulogenic/high fiber diet, which supports the link between a fiber-enriched diet and saccharolytic bacteria. The functional analysis indicated that a flatulogenic diet significantly up-regulated and down-regulated several metabolic pathways. In order to confirm these results, a greater sample size may be required in future studies. Unexpectedly, in contrast to a strict fiber diet, a flatulogenic diet, which increases the volume of intestinal gas in both subjects complaining of excessive gas production (Manichanh *et al.*, 2013), appeared to decrease several categories of functions that are involved in carbohydrate or energy production. Finally, the observed correlation between volume of gas produced and *Bifidobacterium longum* and several functions and categories of functions could be compared in future studies involving strict plant-based or animal-based diet. For future studies, we recommend combining DNA-seq with RNA-seq in order to normalize RNA to DNA (i.e. transcripts per gene) when calculating differential expression between samples. Furthermore, in order to recover both 16S rRNA and mRNA sequences in a non-biased manner and to increase the number of potential mRNA reads at a reasonable cost, we recommend using the same total RNA-extracted sample in two separate experiments: 1) a rRNA removal procedure to enrich mRNA sequences and sequencing at a coverage depth of 10–20 million reads per sample; and 2) a sequencing step with a much lower coverage (100,000 reads per sample) without the rRNA removal step to analyze the active microbial composition in an unbiased manner. We then, could implement, and validate a metatranscriptomic pipeline by making use of the multi-threading capacity of modern computers and then validated its functionality by comparing different methods for taxonomy profiling, by analyzing synthetic mock communities, by analyzing published RNA-seq data and by generating RNA-seq data from fecal samples. The pipeline was implemented on the basis of a constantly changing environment, thus offering the possibility to easily integrate third-party tools, improve parts of the pipeline or change entire modules as long as the input/output folder structure is preserved. The pipeline is available and downloadable from the following webpage: www.metatrans.org, which also provides a tutorial for users.

Along the course of this dissertation, the pipeline has been evolving and adapting to new needs according to software updates and analysis requirements. Though the maintenances of the tool were possible up to some extent, we were able to yet re-validate our tool with a recent published metatatranscriptomics analysis tool from the Huttenhower laboratories, HUMAnN2 (Franzosa *et al.*, 2018). The comparisons showed that regardless of the runtimes, improved by several folds in HUMAnN2, the results we obtained from the evaluation showed very close results in microbiome composition and functional database correlations. Interestingly, we found that using the MetaHIT14 gene catalog we were able to recover more functional annotations than using their default databases. Despite the improvement in the number of aligned reads in HUMAnN2, the number of unknown reads remained high, roughly 30%. Probably, this ratio is still higher since we don't have a current gold standard to assess the quality of assignments. Therefore, it is a matter to set a tradeoff between the sensitivity and specificity by choosing a higher or less similarity cut-off in sequence aligners. In MetaTrans we used higher strict parameters in sequence similarity when mapping reads to databases compared to other pipelines using same sequence aligner (SOAP2, (Li *et al.*, 2009)), like MOCAT (Kultima *et al.*, 2016). Using their aligner parameters, we recovered around 15% more assignments, illustrating the importance of the bias produced when conducting sequence similarities to acquire annotation of reads. Besides, the percentage of unknown functions in protein or gene databases remains quite high between 40% and 60%.

Although HUMAnN2 paper was published in 2018, we noticed that the default UniRef database version used in HUMANN2 was created four years ago (v2014_07), currently, UniRef has a new release (v2019_02) that represents 4-fold the number of annotated proteins compared to the default database. Hence, we recommend updating UniRef to the new version before using the tool. Unfortunately, time restrictions did not allow us to do a re-analysis of all samples with this tool for the time of this writing. Moreover, we detected that the way the feature counts summarization were normalized, did not allow us to use those tables in downstream differential expression analysis tools like DESeq2 (Love, Huber and Anders, 2014) or EdgeR (Robinson, McCarthy and Smyth, 2009), since these tools based on a negative binomial model distribution are better suited for raw counts.

After the development and validation of the pipeline, the metatranscriptomic analysis applied to IBD yield interesting results at RNA level that were comparable to previous publications (Imhann *et al.*, 2018; Lloyd-Price *et al.*, 2019).

Comparisons of healthy relative subjects of CD and UC could not determine any minimal difference between them at any of the timepoints, nor at functional or taxonomical or gene count analysis. These findings suggest that the two groups of healthy controls could be pooled to increase statistical power for further comparisons with each patient group. Comparing healthy controls with IBD patients without separating UC from CD did not show significant differences (at 16S rRNA and mRNA levels), these findings are in agreement with a recent previous work (Lloyd-Price *et al.*, 2019).

Using 16S rRNA data, we identified a significant difference between UC and CD, indicating that these two disease phenotypes presented a different active microbial community composition. In a previous work, using 16S rDNA sequence data on a larger IBD cohort, our group identified a microbial signature for CD (Pascal *et al.*, 2017). Our findings, at the rRNA level, could detect part of this microbial signature at baseline and the last timepoint, where half of the patients evolved to relapse. For instance, our rRNA analysis pointed out that the *Fusobacterium* genus, also one of the most relevant genus found in the Crohn's signature at the DNA level, was one of the most over-expressed (up to 5-6 log2FC times higher) in CD patients either at baseline or at the last timepoint in comparison with UC. Furthermore, this genus was strongly correlated (rho=0.8, FDR<0.05) with the CARD9.rs4077515 gene, which was found associated with CD and UC in a previous publication (Zhernakova *et al.*, 2008). Interestingly, the

CARD9 gene was identified as an intestinal epithelial cell restituent in mice (Sokol *et al.*, 2013), but not its SNP variant CARD9.rs4077515, which behaves as pro-inflammatory, and was established as a risk factor in the development of ileal CD (Zhong *et al.*, 2018, 2019) .

Three of the lower abundant genera identified at DNA level in the Crohn's microbial signature, were also relevant at RNA level: *Anaerostipes* (a butyrate producer, only detected at basal comparison), *Methanobrevibacter* (an obligate anaerobe methane producer), *Faecalibacterium prausnitzii* (a butyrate producer, critical short chain fatty acid in maintaining homeostasis in the colon) and members of the family Christensenellaceae. Interestingly, the last three genera were found only at the last time point, when the disease was more severe, suggesting that those actors might play an important role in disease severity.

Additionally, *Ruminococus gnavus*, a key actor in both UC and CD as described in previous studies (Henke *et al.*, 2019; Lloyd-Price *et al.*, 2019; Yilmaz *et al.*, 2019), was found, in our study using 16S rRNA data, in higher abundance only in CD patients (3.6 log2FC times) at both baseline and last-time point, but not in UC patients. Yilmaz et al. also pointed *Blautia* and *Faecalibacterium* as key players in IBD. We found two members of Blautia differentially expressed in our analyses: *Blautia producta* was significantly upregulated at baseline, whereas *Blautia obeum* was found downregulated in the last timepoint. It is also worth mentioning that we observed a strong positive association between *Ruminococcus* and *Akkermansia*, found decreased many fold in CD and UC (Png *et al.*, 2010).

Two bacterial groups, at the family level, Paraprevotellaceae and Peptococcaceae, were found strongly downregulated in CD, at baseline and at the last time-point, respectively. Peptococcaceae was also negatively correlated with IL-1β, a proinflammatory cytokine (Regner *et al.*, 2018). Intriguingly, a comparison with samples metadata uncovered that only all non-smokers and one ex-smoker UC patients had abundance of Peptococcaceae, and the patient with the highest abundance had the longest duration of the disease (23 years), and evolved to relapse during the course of the follow-up. Finally, Peptococcaceae as well as Christensenellaceae, were significantly decreased (FDR<0.05) in a broad study comparing 582 healthy controls and 313 patients with clinical phenotype of IBD (Imhann *et al.*, 2018).

The alpha-diversity analysis (Shannon index) in UC and CD patients based 16S rRNA data, showed concordance with the 16S rDNA analyses performed in previous literature, where bacterial diversity was significantly lower in CD. This finding validated, at 16S rRNA level, previous discoveries regarding the loss of active microbial species in CD compared to non-CD subjects.

The functional analysis of the putative mRNA from expressed transcripts, showed higher differences between healthy relatives (HR) of CD and UC-CD than differences between HR-UC. The latter only exhibited differences when using the unsupervised functional database EggNOG. The annotation of genes at functional categories allowed us to identify four major functions capturing the higher differential expression (log2FC>1, FDR<0.05) in UC: Cell motility ([N]; downregulated), Cell cycle control, cell division, chromosome partitioning ([D]; upregulated), Translation, ribosomal structure and biogenesis ([J]; upregulated) and unknown functions (downregulated). These functions suggest a reduction in cell motility, while cell replication and protein synthesis activity were being carried out. Oddly, functions related to carbohydrate metabolism ([G]), known to produce short chain fatty acids (SCFA) (Venegas *et al.*, 2019), were not reduced (Lloyd-Price *et al.*, 2019; Venegas *et al.*, 2019) but slightly increased (log2FC=0.5). Altogether, these findings might emphasize that differences between UC and healthy subjects are weak.

A more significant and consistent functional difference was observed when analyzing patients with CD versus non-CD individuals. Notably, in the comparison with healthy relatives, we identified 10 out of 33 pathways differentially expressed (p-value<0.05), in agreement with previous findings (Imhann *et al.*, 2018), were the authors performed similar comparisons using a much larger cohort of patients and 16SrDNA sequence data for bacterial detection and functional prediction. The main relevant encoding functions were related to Metabolism and Cell motility, though two of them were related to Genetic information processing (tRNA and DNA replication). Interestingly, one pathway associated with the production of short-chain fatty acids (map00540; propanoate or propionate, contributes to glucose synthesis) was found significantly decreased. SCFA are crucial metabolites for the digestion and homeostasis of the gastrointestinal tract, and are known to have protective effects on intestinal barrier (Scheppach, 1994; Vinolo *et al.*, 2011; Liu *et al.*, 2012; Feng *et al.*, 2018; Venegas *et al.*, 2019). However, levels of Lipopolysaccharides (LPS), known to be inhibited by SCFA (Li *et al.*, 2018) were downregulated, probably due to the remission status of CD patients. On the other hand, Peptidoglycan pathway (map00550) was highly enriched with upregulated genes in CD patients. It is known to be the target of antibiotics like β-lactam, as is critical in cell structure. Of note, methane metabolism pathway enrichment was diminished in CD patients (He *et al.*, 2017), related with energy production, as well the two most significant underexpressed pathways related with cell motility. By contrast, the Glutathione metabolism pathway was found uniquely enriched in upregulated CD genes, found also in a metagenomic study as exhibiting enhanced potential for antioxidant defense (He *et al.*, 2017).

Finally, in the last comparison, contrasting patients of UC and CD, we identified 22 out of 28 detected enriched pathways (differentially expressed genes) as significant and shared with the previous comparison with healthy relatives. Interestingly, only 4 of the shared set were found in different up/down regulation, whereas, in general, highest differences in expression were displayed in CD. This suggests that functional differences in UC do not differ greatly from healthy subjects, supporting the hypothesis of Pascal et al. (Pascal *et al.*, 2017). Among the four pathways that mostly differ, Pyruvvate metabolism (map00620) was found in greatly downregulated in CD when compared to healthy relatives. In contrast Carbon metabolism (map01200) was found strongly upregulated when CD was compared to UC. The alpha diversity, using functional data, was found always higher in CD either comparing with UC or healthy. This finding is contradictory to the results found at the taxonomic level, where healthy subjects account for a greater compositional diversity.

Overall, we found more differences with patients at basal timepoint (REM; n=14) than at last timepoint (REM/REL; n=7) quite probably due to small sample size. Any attempt to make comparisons with the severity of the disease (REL, n=7) was unsuccessful. If differences between groups are not very strong, the power to detect significances is very low with this small cohort. The only way to overcome this lack of power effectively relies on trying to obtain higher number of samples per condition. Despite this, we could observe that differences between HR and CD were more significant than between HR and UC, probably highlighting that UC is less dysbiotic than CD compared to HR at the active microbiome level, which is also in agreement with previous findings (Lloyd-Price *et al.*, 2019). A prediction of relapse (REL) through the analysis of samples collected at the time of remission (REM) using machine learning techniques as, for instance, random forest, was not possible as we did not obtain enough statistical power to detect significant differences between groups.

We have discussed the relevant microbial organisms and functional features according to our findings. Nonetheless, all results, discussions and conclusions must be put in the context of the limitation of this study, and thus taken with caution. The cost of sequencing metatranscriptomic samples were around one thousand and five hundred euros at the time we started the pilot study. This cost is much higher than for the 16S rDNA analysis that may require less than 150 euros per sample. Difficulties during RNA extraction are well known, as well as its limited lifetime before degradation. Sequencing machines also add error rates in sequenced reads that must be treated properly with bioinformatic tools to minimize the risk of low-quality bases. Downstream analyses are also subject to constant software updates and different methods of analysis, and still lack of a widely settled consensus in methodologies to make other works easily comparable, though effort is put in that direction with reproducible research. Further, still a high and notorious 50% of functional diversity is still to be characterized and linked to taxa. The MetaHIT14 gene catalog, for instance, only has annotated 14% of the genes at Genus rank (roughly 20% at Phylum level). The new version of the unsupervised functional database EggNOG is now getting better annotations, but the old version (like EggNOG3.0) still lacked external and more comprehensible functional annotations. We encountered more difficulties interpreting results at the functional level using unsupervised databases than supervised databases though at risk of having much less annotation rates. Validation of these pipelines are not a minor complication either, because of the complexity to simulate a mock community that fairly resembles to real microbiome compositions, or the limited use of simulated mock communities. Finally, metatranscriptomics gives a single snapshot in a particular moment, when it is, in fact, a changing and complex interacting environment. After all, if all this is taken into consideration, and without new advanced techniques, this work provides a better understanding of the functional mechanisms characterizing the active gut microbiome.

Chapter 6.

Conclusions

6 Conclusions

The results obtained in this dissertation allowed us to draw the following conclusions:

1. We developed a metatranscriptomic analysis tool that performs functional and taxonomical analysis, taking advantage of multithreading architecture of computers. This tool processes samples in a reasonable time and generates results comparable to current rRNA analysis tools.

2. The development of this bioinformatics pipeline has allowed us to understand our limitations in the metaomics field as well as to gain insights into the characterization of the active human gut microbiome.

3. Gene expression analysis validates some hypotheses previously proposed at the genomic level. However, taxonomic profiles obtained from 16S rRNA data contrasted with those obtained with 16S rDNA data, indicating that not all microorganisms identified at the DNA level play an active role in both IBS and IBD.

4. Our study on IBD cohort revealed that CD and UC presented a distinct active microbiome profile at the taxonomic as well as functional level. Furthermore, CD patients showed greater dysbiosis than UC patients. Altogether, these results validate previous works based on gene content analysis.

5. Our results at RNA level also suggested that dysregulations of different pathways related to the Short Chain Fatty Acids metabolism and cell survival were associated with disease severity.

6. Finally, our study provides a very comprehensive description of the active microbial functions and pave the way for future investigations on inflammatory bowel diseases.

Chapter 7.

Future lines

7 Future lines

At this point, it seems clear that metatranscriptomics is able to unravel insights into the most important microbial activities in human diseases such as IBS or IBD, located in the gastrointestinal tract with highest bacterial concentrations (Sender, Fuchs and Milo, 2016). As in metagenomics, in metatranscriptomics analysis functional genes are still poorly annotated, and we must be cautious about the interpretation of the results. At the time of this writing, we are still far from knowing with exactitude the number of functions the gut microbiome encodes or expresses. In the last release of the human gene catalog (Li *et al.*, 2014) roughly fifty percent of the genes had unknown annotated functions. This highlights how important still is the unknown category when displaying functional compositions. They may be other between-group differences that are hidden in the proportion of unknown genes, which might reveal themselves in future re-analyses, strengthen by improved gene annotation. One of the most widely used and curated metabolic pathways database, KEGG (Kanehisa *et al.*, 2017), became a paid-resource some years ago for academic use, and new release are no longer free of use. Now, other academic-free data bases like MetaCyc (Caspi *et al.*, 2018), using a curated multiorganism pathway database, are increasing their popularity and it is advisable to use this database in future analysis if paid-resources are not affordable. However, the use of updated or different pathway databases in the human gut gene catalog will require an extra effort to re-annotate all genes in the catalog.

The RNA abundance normalization through DNA abundance per sample is also an important factor to achieve fair results when comparing samples. Yet this step requires higher sequencing costs and computational resources. A new manuscript focusing on this methodology is under preparation by our group. Furthermore, an increase in biological experimental replications is also an important issue to properly identify differentially expressed genes, as all statistical models underlying these tools rely on biological replicates (Schurch *et al.*, 2016) to estimate regulated genes more accurately.

The continuous progress in this area of research will require the incorporation of more sophisticated and faster bioinformatics tools in the analysis. Therefore, the maintenance and the improvement of those tools is a desirable feature to keep up to date future analysis with new annotations and improved algorithms.

As it has been commented in the discussion chapter, the increase in sample size is mandatory to increase significant levels and to sharpen results, even though we are aware of the inherent difficulties to address this issue.

All the results obtained in this work serve as a basis of comparison to identify new targets of functions or taxa that can eventually be associated with IBD or IBS in future analyses.

Chapter 8.

Appendices

8 Appendices

8.1 Appendix A - Supporting material for Chapter 4.

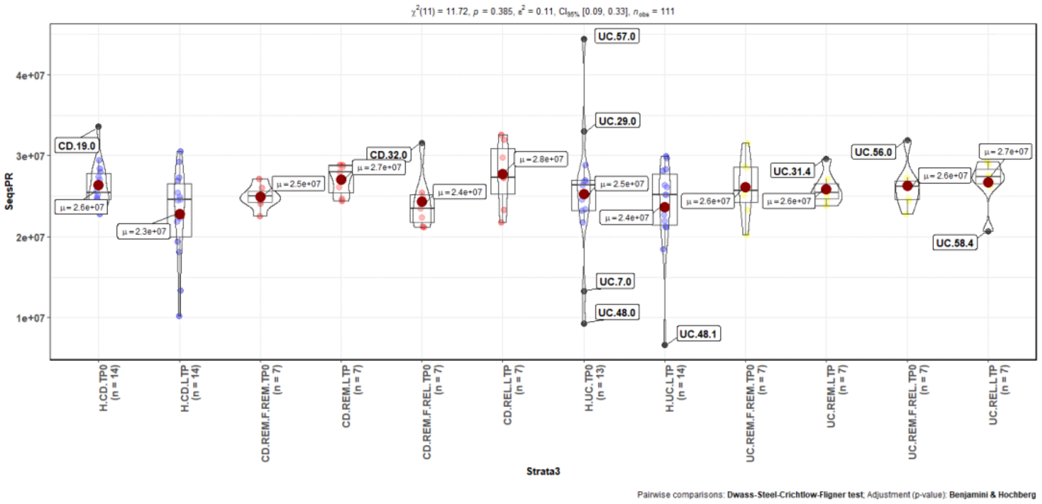


Figure A 1 Sequencing depth (paired-reads) on Illumina Hi-Seq 2000
Distribution of the number of paired-end sequencing reads grouped by healthy (H - blue color), and patients of CD (red) and UC (yellow). Groups are further classified by their basal (timepoint0) or last timepoint (LTP), and by their patient status (REM – remission, REL – relapse). Within the basal timepoint the groups of REM are differentiated between those that remain in REM state in the last timepoint (F.REM, where F stands for future) and those that fall into REL state (F.REL). Each distribution is characterized by a boxplot in the middle and their density distribution along both sides. Median is depicted by a black band inside the box, and the mean (µ) by a red dot. Outliers are labeled by their corresponding sample ids. Stats on the top of the chart show stats relative to the comparison of all groups (p-value = 0.385, Kruskal-Wallis). Stats on the bottom refer to a post-hoc pairwise analysis using FDR correction (no statistically significance between groups were found).

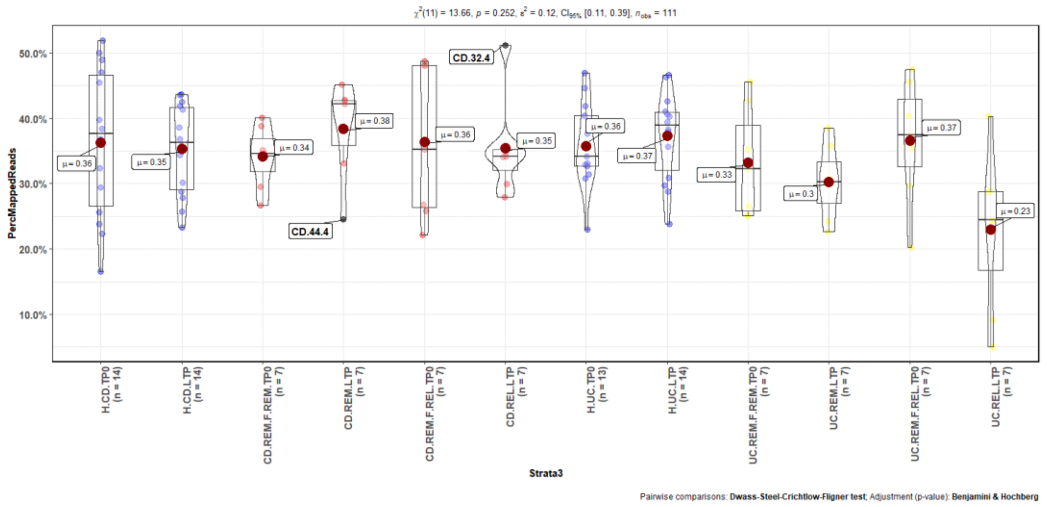
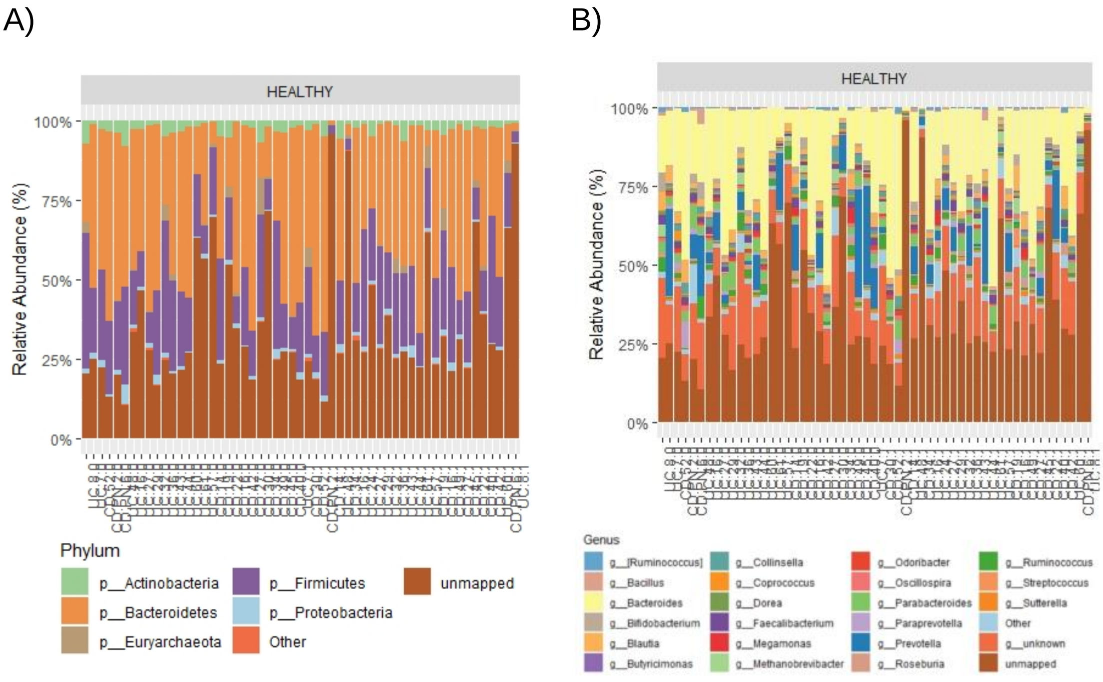
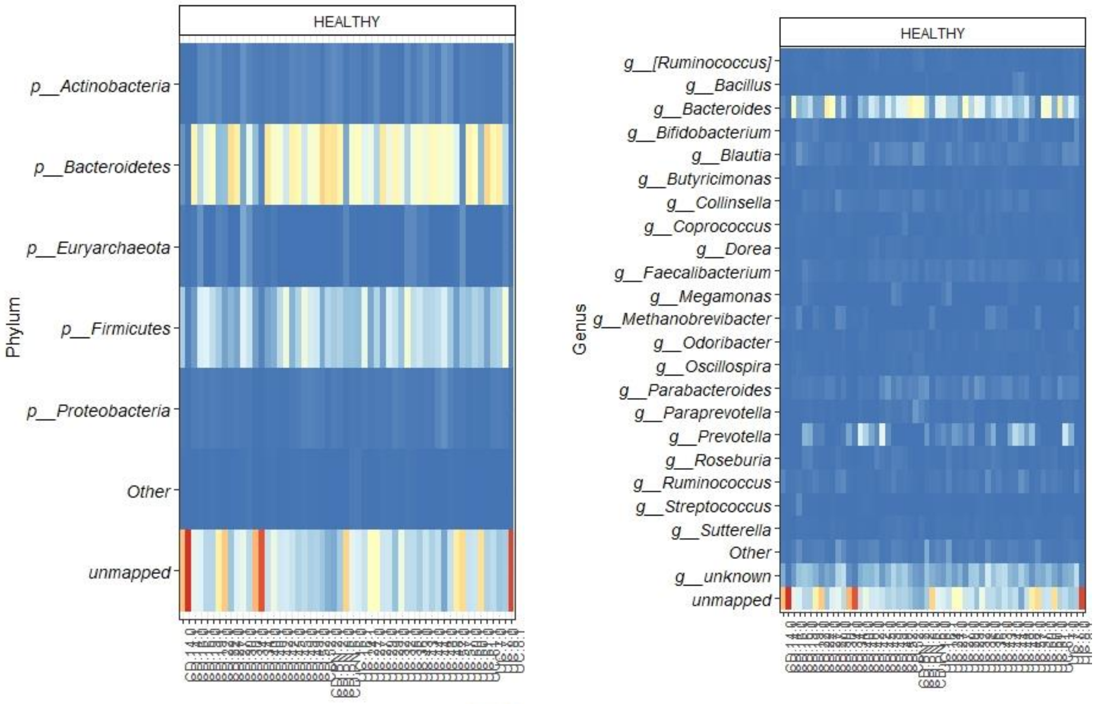


Figure A 2 Percentage of reads mapped to the MetaHIT-14 gene catalog.
Distribution of the percentage of reads that were mapped to the MetaHIT-14 gene catalog grouped by healthy (H - blue color), and patients of CD (red) and UC (yellow). Groups are further classified by their basal (timepoint0) or last timepoint (LTP), and by their patient status (REM – remission, REL – relapse). Within the basal timepoint the groups of REM are differentiated between those that remain in REM state in the last timepoint (F.REM, where F stands for future) and those that fall into REL state (F.REL). Each distribution is characterized by a boxplot in the middle and their density distribution along both sides. Median is depicted by a black band inside the box, and the mean (μ) by a red dot. Outliers are labeled by their corresponding sample ids. Stats on the top of the chart show stats relative to the comparison of all groups (p=0.385, Kruskal-Wallis). Stats on the bottom refer to a post-hoc pairwise analysis using FDR correction (no statistically significance between groups were found).

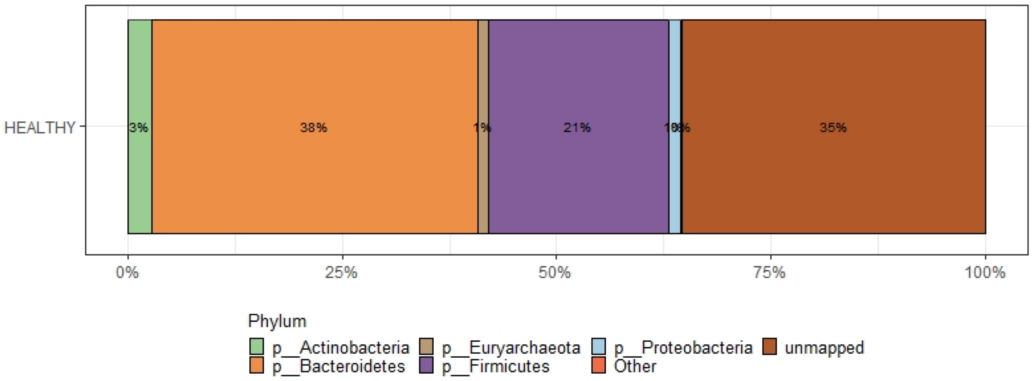


C)

D)


HEALTHY
Phylum
p__Actinobacteria
p__Bacteroidetes
p__Euryarchaeota
p__Firmicutes
p__Proteobacteria
Other
unmapped
HEALTHY
Genus
g__[Ruminococcus]
g__Bacillus
g__Bacteroides
g__Bifidobacterium
g__Blautia
g__Butyricimonas
g__Collinsella
g__Coprococcus
g__Dorea
g__Faecalibacterium
g__Megamonas
g__Methanobrevibacter
g__Odoribacter
g__Oscillospira
g__Parabacteroides
g__Paraprevotella
g__Prevotella
g__Roseburia
g__Ruminococcus
g__Streptococcus
g__Sutterella
Other
g__unknown
unmapped

E)


HEALTHY
3%
38%
1%
21%
1%
0%
35%
0%
25%
50%
75%
100%
Phylum
p__Actinobacteria
p__Bacteroidetes
p__Euryarchaeota
p__Firmicutes
p__Proteobacteria
Other
unmapped

F)

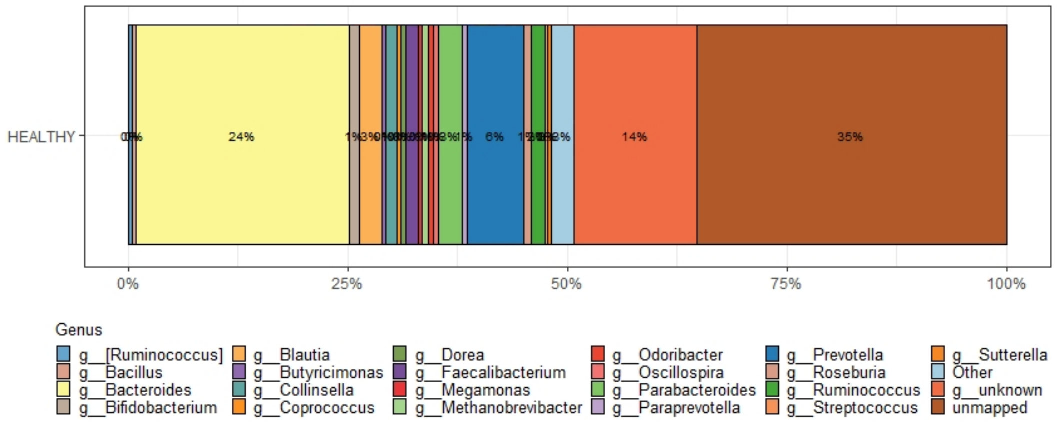
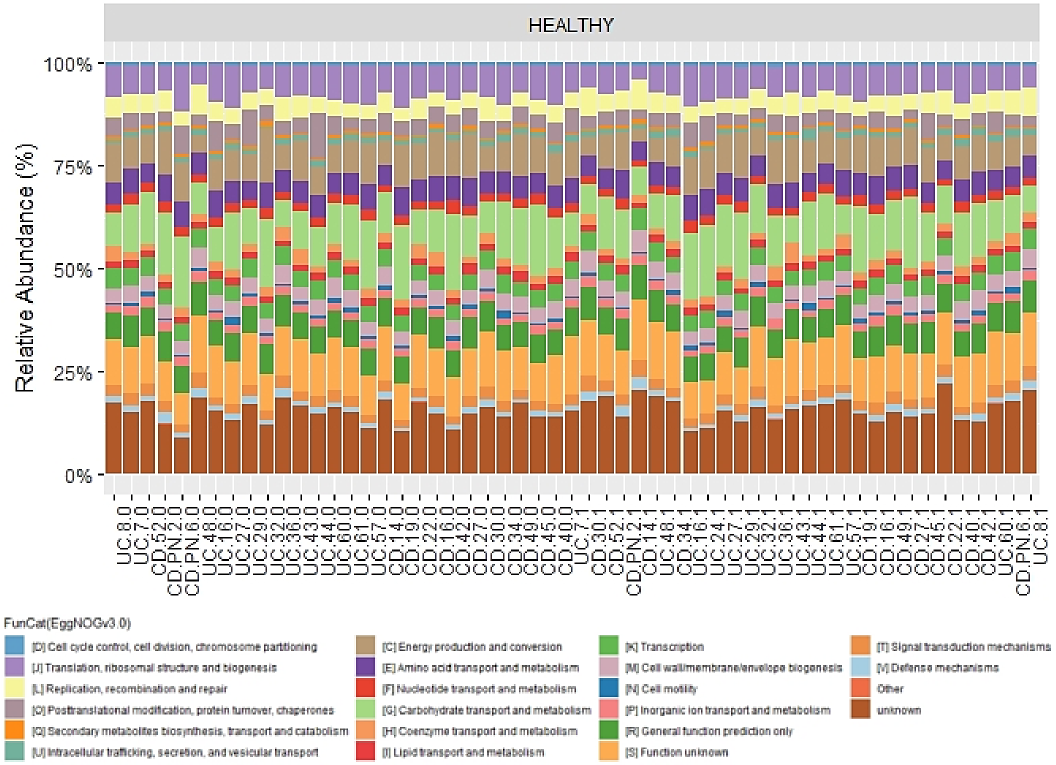


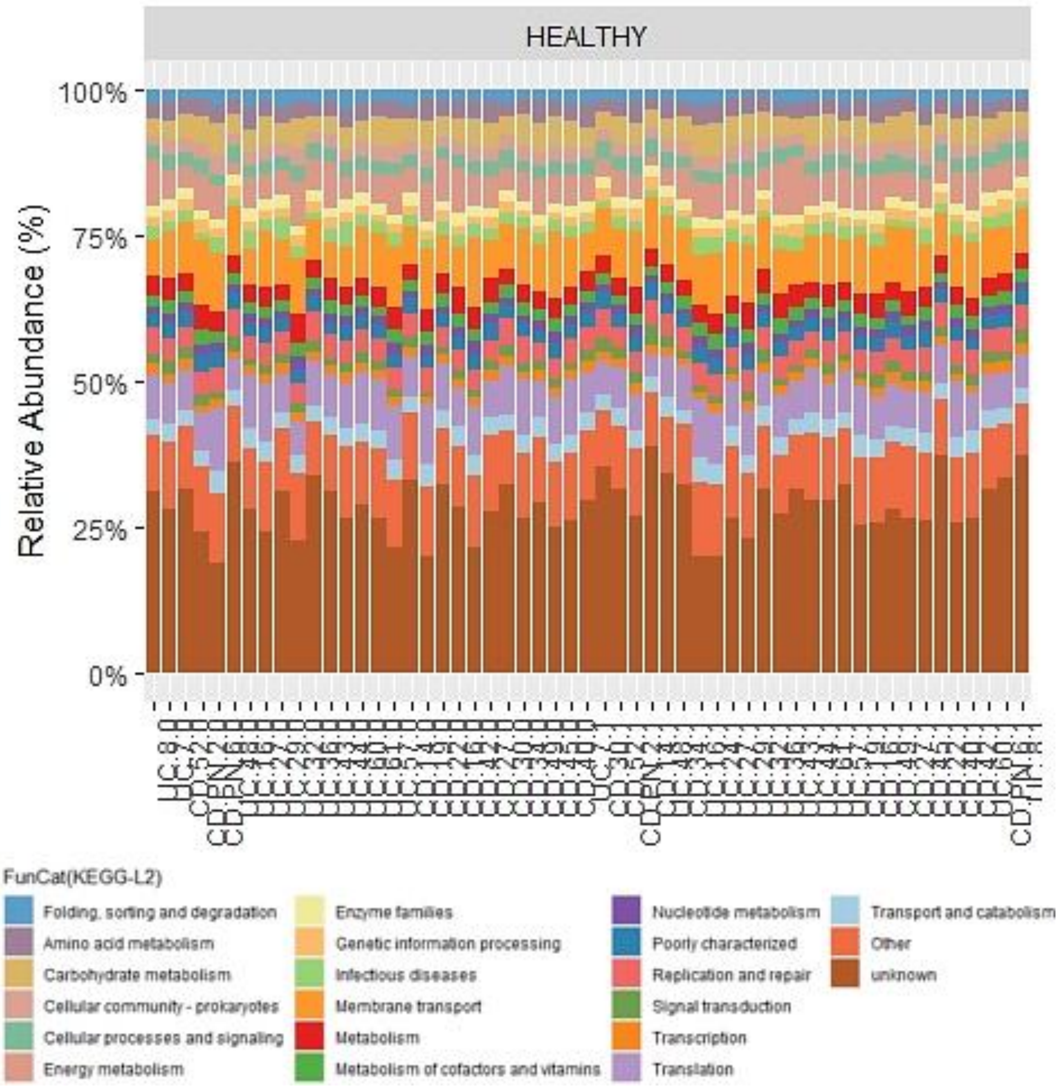
Figure A 3 Taxonomical composition of the HEALTHY active microbiome

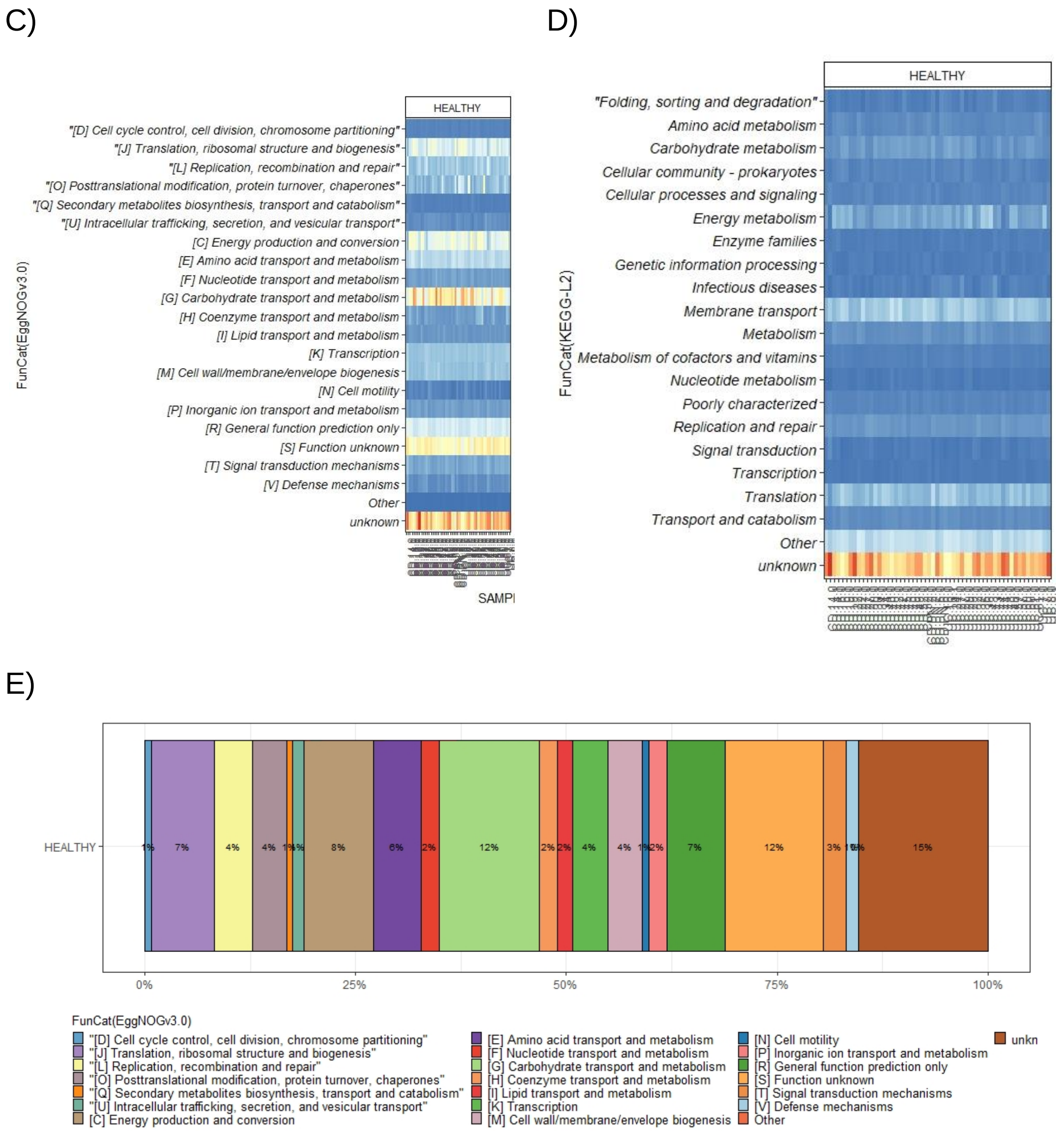
All organisms for which their function is unknown are collapsed to the “unknown” category. The remaining taxons below 3% cutoff are collapsed to “Other” to improve an overall overview of the composition. A) and B) barplots, show the individual relative abundance composition of healthy subjects for the two taxonomical annotation ranks, Phylum and Genus. C) and D) display the relative abundance in a heatmap chart to better appreciate differences in abundances between samples and taxa. E) and F) show the average proportions of taxonomical relative abundances of healthy subjects

A)


HEALTHY
100%
75%
50%
25%
0%
Relative Abundance (%)
UC.8.0
UC.7.0
CD.52.0
CD.PN.2.0
CD.PN.6.0
UC.48.0
UC.16.0
UC.27.0
UC.29.0
UC.32.0
UC.36.0
UC.43.0
UC.44.0
UC.60.0
UC.61.0
UC.57.0
CD.14.0
CD.19.0
CD.22.0
CD.16.0
CD.42.0
CD.27.0
CD.30.0
CD.34.0
CD.49.0
CD.45.0
CD.40.0
UC.7.1
CD.30.1
CD.52.1
CD.PN.2.1
CD.14.1
UC.48.1
CD.34.1
UC.16.1
UC.24.1
UC.27.1
UC.29.1
UC.32.1
UC.36.1
UC.43.1
UC.44.1
UC.61.1
UC.57.1
CD.19.1
CD.16.1
CD.49.1
CD.27.1
CD.45.1
CD.22.1
CD.40.1
CD.42.1
UC.60.1
CD.PN.6.1
UC.8.1
FunCat(EggNOGv3.0)
[D] Cell cycle control, cell division, chromosome partitioning
[J] Translation, ribosomal structure and biogenesis
[L] Replication, recombination and repair
[O] Posttranslational modification, protein turnover, chaperones
[Q] Secondary metabolites biosynthesis, transport and catabolism
[U] Intracellular trafficking, secretion, and vesicular transport
[C] Energy production and conversion
[E] Amino acid transport and metabolism
[F] Nucleotide transport and metabolism
[G] Carbohydrate transport and metabolism
[H] Coenzyme transport and metabolism
[I] Lipid transport and metabolism
[K] Transcription
[M] Cell wall/membrane/envelope biogenesis
[N] Cell motility
[P] Inorganic ion transport and metabolism
[R] General function prediction only
[S] Function unknown
[T] Signal transduction mechanisms
[V] Defense mechanisms
Other
unknown

B)


HEALTHY
Relative Abundance (%)
100%
75%
50%
25%
0%
FunCat(KEGG-L2)
Folding, sorting and degradation
Amino acid metabolism
Carbohydrate metabolism
Cellular community - prokaryotes
Cellular processes and signaling
Energy metabolism
Enzyme families
Genetic information processing
Infectious diseases
Membrane transport
Metabolism
Metabolism of cofactors and vitamins
Nucleotide metabolism
Poorly characterized
Replication and repair
Signal transduction
Transcription
Translation
Transport and catabolism
Other
unknown


C)
HEALTHY
"[D] Cell cycle control, cell division, chromosome partitioning"
"[J] Translation, ribosomal structure and biogenesis"
"[L] Replication, recombination and repair"
"[O] Posttranslational modification, protein turnover, chaperones"
"[Q] Secondary metabolites biosynthesis, transport and catabolism"
"[U] Intracellular trafficking, secretion, and vesicular transport"
[C] Energy production and conversion
[E] Amino acid transport and metabolism
[F] Nucleotide transport and metabolism
[G] Carbohydrate transport and metabolism
[H] Coenzyme transport and metabolism
[I] Lipid transport and metabolism
[K] Transcription
[M] Cell wall/membrane/envelope biogenesis
[N] Cell motility
[P] Inorganic ion transport and metabolism
[R] General function prediction only
[S] Function unknown
[T] Signal transduction mechanisms
[V] Defense mechanisms
Other
unknown
FunCat(EggNOGv3.0)
SAMPI
D)
HEALTHY
"Folding, sorting and degradation"
Amino acid metabolism
Carbohydrate metabolism
Cellular community - prokaryotes
Cellular processes and signaling
Energy metabolism
Enzyme families
Genetic information processing
Infectious diseases
Membrane transport
Metabolism
Metabolism of cofactors and vitamins
Nucleotide metabolism
Poorly characterized
Replication and repair
Signal transduction
Transcription
Translation
Transport and catabolism
Other
unknown
FunCat(KEGG-L2)
E)
HEALTHY
1%
7%
4%
4%
1%
1%
8%
6%
2%
12%
2%
2%
4%
4%
1%
2%
7%
12%
3%
1%
1%
15%
0%
25%
50%
75%
100%
FunCat(EggNOGv3.0)
"[D] Cell cycle control, cell division, chromosome partitioning"
"[J] Translation, ribosomal structure and biogenesis"
"[L] Replication, recombination and repair"
"[O] Posttranslational modification, protein turnover, chaperones"
"[Q] Secondary metabolites biosynthesis, transport and catabolism"
"[U] Intracellular trafficking, secretion, and vesicular transport"
[C] Energy production and conversion
[E] Amino acid transport and metabolism
[F] Nucleotide transport and metabolism
[G] Carbohydrate transport and metabolism
[H] Coenzyme transport and metabolism
[I] Lipid transport and metabolism
[K] Transcription
[M] Cell wall/membrane/envelope biogenesis
[N] Cell motility
[P] Inorganic ion transport and metabolism
[R] General function prediction only
[S] Function unknown
[T] Signal transduction mechanisms
[V] Defense mechanisms
Other
unkn

F)

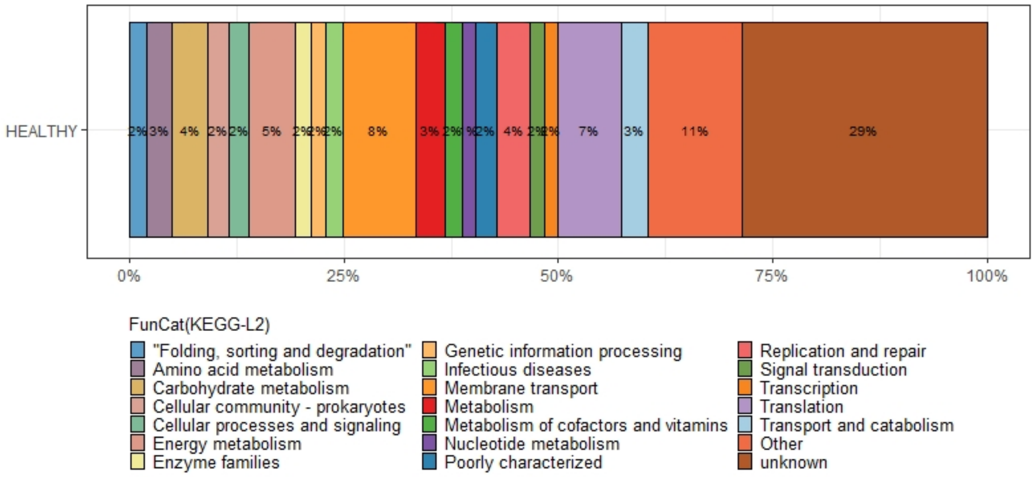
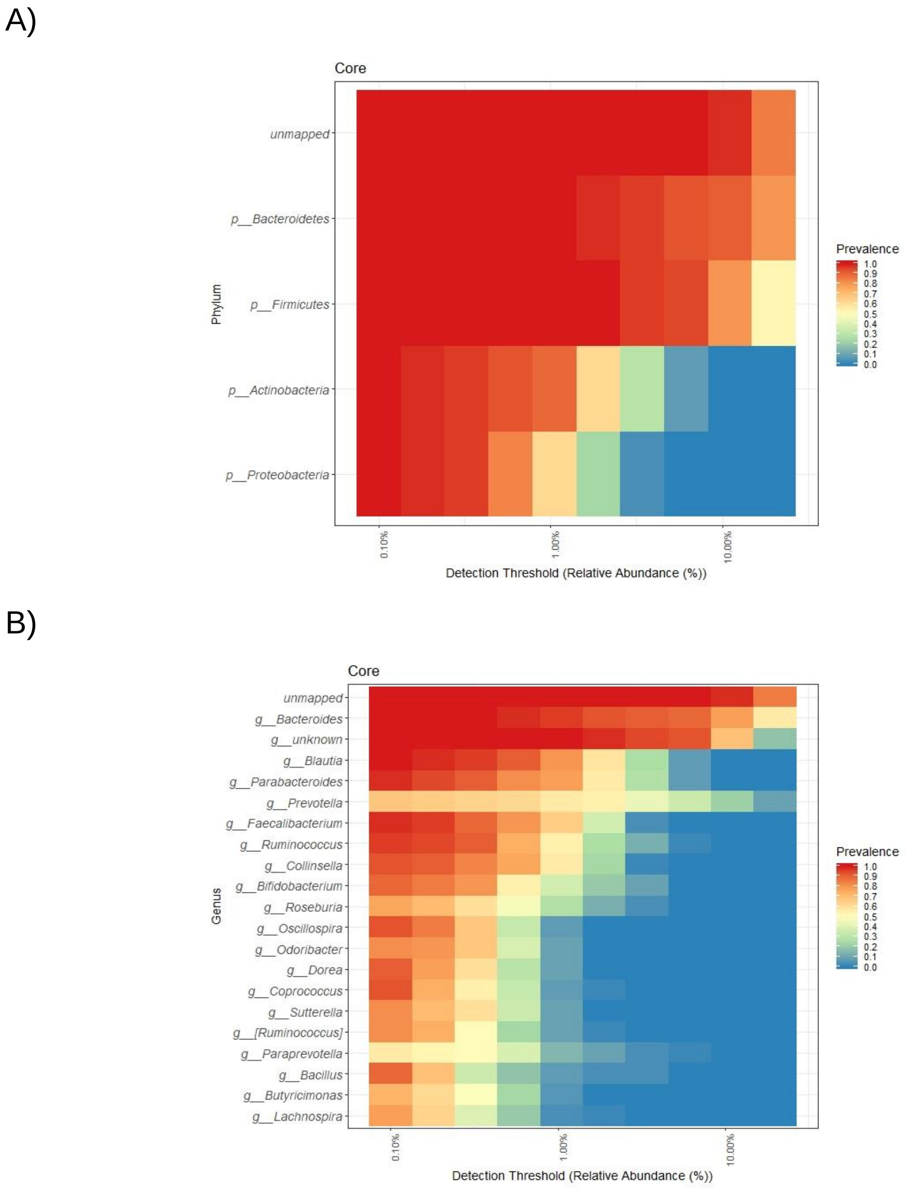


Figure A 4 Functional composition of the HEALTHY active microbiome

All organisms for which their function is unknown are collapsed to the “unknown” category. The remaining functions below 3% cutoff are collapsed to “Other” to improve an overall overview of the composition. A) and B) barplots, show the individual relative abundance composition of healthy subjects for the two functional annotation databases, EggNOGv3 and KEGG, at their functional categories annotation level (KEGG at Level2). C) and D) display the relative abundance in a heatmap chart to better appreciate differences in abundances between samples and types of functions. E) and F) show the average proportions of functional relative abundances of healthy subjects.


A)
Core
unmapped
p__Bacteroidetes
p__Firmicutes
p__Actinobacteria
p__Proteobacteria
Phylum
Prevalence
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.10%
1.00%
10.00%
Detection Threshold (Relative Abundance (%))
B)
Core
unmapped
g__Bacteroides
g__unknown
g__Blautia
g__Parabacteroides
g__Prevotella
g__Faecalibacterium
g__Ruminococcus
g__Collinsella
g__Bifidobacterium
g__Roseburia
g__Oscillospira
g__Odoribacter
g__Dorea
g__Coprococcus
g__Sutterella
g__[Ruminococcus]
g__Paraprevotella
g__Bacillus
g__Butyricimonas
g__Lachnospira
Genus
Prevalence
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.10%
1.00%
10.00%
Detection Threshold (Relative Abundance (%))

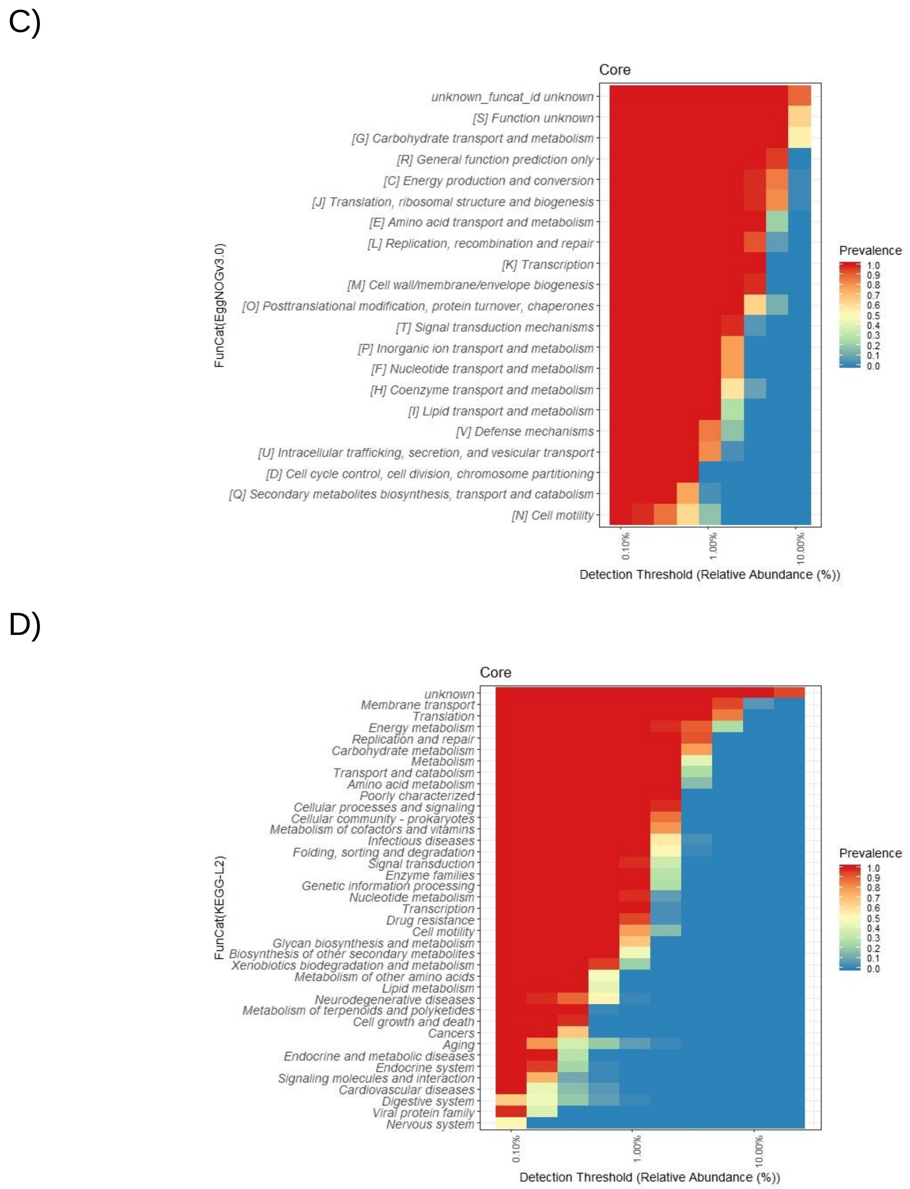


Figure A 5 Healthy microbiome core

Core heatmaps displaying prevalence of features (taxa/functional) at different relative abundance detection cutoffs (>0.1 - >10%). Prevalence is measured as the rate of relative abundance selected among all samples in a 0-1 scale, where 1 represents a 100% prevalence of the feature along all samples. Features shown are those identified at a minimum detection of 0.1% of relative abundance and a minimum prevalence of 50%. A) and B) display the core microbiome of healthy subjects in the taxonomical analysis at their Phylum and Genus rank respectively. C) and D) display the microbiome of healthy subjects in the functional analysis using the EggNOGv3 non-supervised functional database and the curated KEGG database at level2.

Table A 1 Percentage of top10 most abundant and prevalent taxa at phylum rank

Top10 most abundant taxa	Percentage(%)	Most prevalent taxa	Percentage(%)
p__Bacteroidetes	37.86	**p__Bacteroidetes**	100
unmapped	35.72	**p__Firmicutes**	100
p__Firmicutes	20.8	**p__Actinobacteria**	100
p__Actinobacteria	2.78	**p__Proteobacteria**	100

p__Euryarchaeota	1.42	unmapped	100
p__Proteobacteria	1.19		
p__[Thermi]	0.06		
p__Acidobacteria	0.05		
p__Armatimonadetes	0.05		
p__Chlamydiae	0.04		

Table A 2 Percentage of top10 most abundant and prevalent taxa at genus rank

Top10 most abundant taxa	Percentage(%)	Most prevalent taxa	Percentage(%)
unmapped	35.72	g__Bacteroides	100
g__Bacteroides	23.93	g__Blautia	100
g__unknown	13.81	unmapped	100
g__Prevotella	6.36	g__unknown	100
g__Parabacteroides	2.67	g__Parabacteroides	98.25
g__Blautia	2.58	g__Faecalibacterium	98.25
g__Ruminococcus	1.58	g__Ruminococcus	96.49
g__Faecalibacterium	1.55	g__Collinsella	92.98
g__Collinsella	1.25	g__Oscillospira	92.98
g__Bifidobacterium	1.19	g__Dorea	91.23
		g__Bifidobacterium	89.47
		g__Bacillus	89.47
		g__Coprococcus	89.47
		g__[Ruminococcus]	82.46
		g__Odoribacter	82.46
		g__Sutterella	82.46
		g__Lachnospira	78.95
		g__Roseburia	77.19

Table A 3 Percentage of top10 most abundant and prevalent functions at functional categories (EggNOGv3.0) annotation level

Top10 most abundant taxa	Percent age(%)	Most prevalent taxa	Percent age(%)
unknown_funcat_id unknown	15.31	unknown_funcat_id unknown	100
[G] Carbohydrate transport and metabolism	11.75	[G] Carbohydrate transport and metabolism	100
[S] Function unknown	11.63	[S] Function unknown	100
[C] Energy production and conversion	8.2	[C] Energy production and conversion	100
[J] Translation, ribosomal structure and biogenesis	7.43	[J] Translation, ribosomal structure and biogenesis	100
[R] General function prediction only	6.96	[R] General function prediction only	100

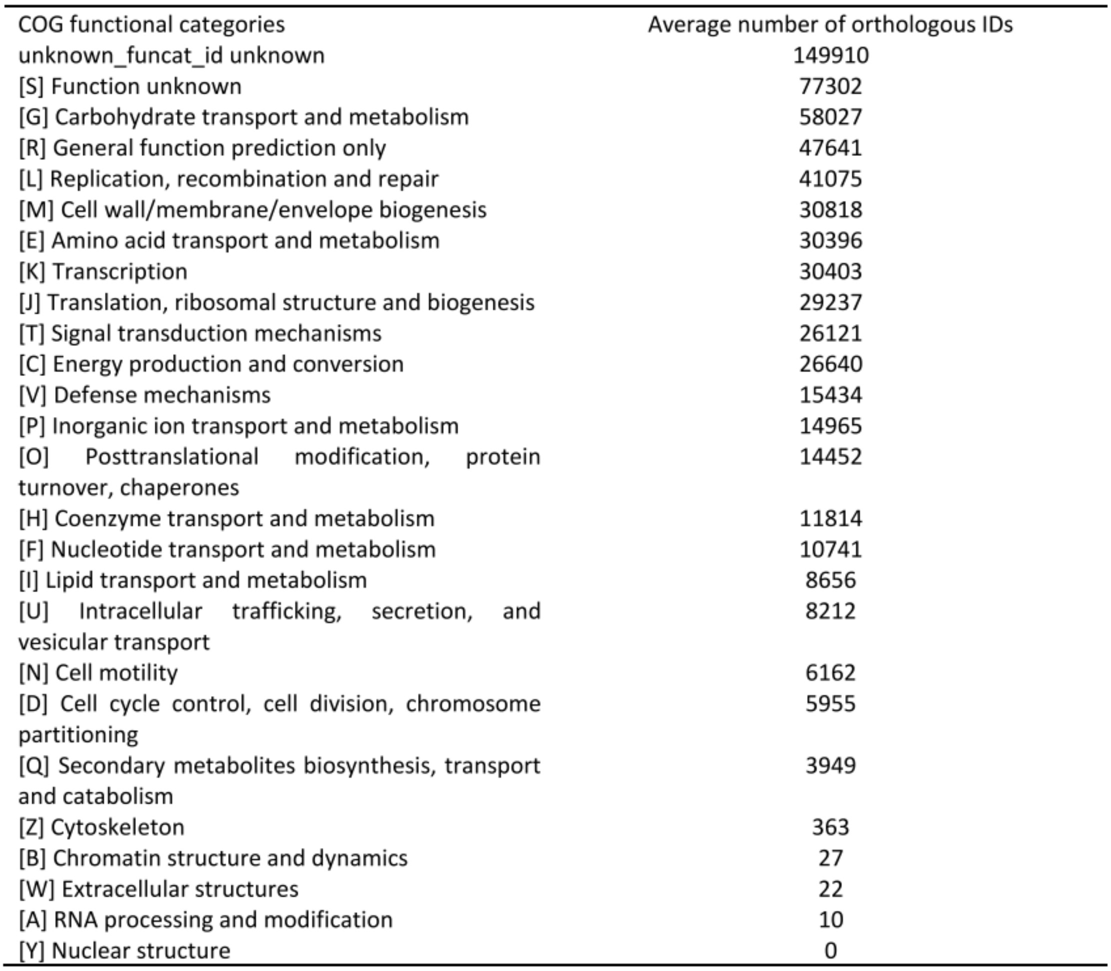
[E] Amino acid transport and metabolism	5.65	[E] Amino acid transport and metabolism	100
[L] Replication, recombination and repair	4.51	[L] Replication, recombination and repair	100
[K] Transcription	4.2	[K] Transcription	100
[O] Posttranslational modification, protein turnover, chaperones	4.16	[O] Posttranslational modification, protein turnover, chaperones	100
		[M] Cell wall/membrane/envelope biogenesis	100
		[T] Signal transduction mechanisms	100
		[F] Nucleotide transport and metabolism	100
		[P] Inorganic ion transport and metabolism	100
		[H] Coenzyme transport and metabolism	100
		[I] Lipid transport and metabolism	100
		[V] Defense mechanisms	100
		[U] Intracellular trafficking, secretion, and vesicular transport	100
		[D] Cell cycle control, cell division, chromosome partitioning	100
		[Q] Secondary metabolites biosynthesis, transport and catabolism	100
		[N] Cell motility	100

Table A 4 Percentage of top10 most abundant and prevalent functions at KEGG-L2 functional categories annotation level

Top10 most abundant taxa	Percentage (%)	Most prevalent taxa	Percentage (%)
unknown	28.56	unknown	100
Membrane transport	8.42	Membrane transport	100
Translation	7.47	Translation	100
Energy metabolism	5.3	Energy metabolism	100
Carbohydrate metabolism	4.12	Carbohydrate metabolism	100
Replication and repair	3.95	Replication and repair	100
Metabolism	3.43	Metabolism	100
Transport and catabolism	3.12	Transport and catabolism	100
Amino acid metabolism	2.97	Amino acid metabolism	100
Cellular processes and signaling	2.46	Cellular processes and signaling	100

	Poorly characterized	100
	Cellular community - prokaryotes	100
	Metabolism of cofactors and vitamins	100
	Infectious diseases	100
	Folding, sorting and degradation	100
	Enzyme families	100
	Nucleotide metabolism	100
	Genetic information processing	100
	Signal transduction	100
	Transcription	100
	Drug resistance	100
	Cell motility	100
	Glycan biosynthesis and metabolism	100
	Biosynthesis of other secondary metabolites	100
	Xenobiotics biodegradation and metabolism	100
	Metabolism of other amino acids	100
	Neurodegenerative diseases	100
	Lipid metabolism	100
	Metabolism of terpenoids and polyketides	100
	Cell growth and death	100
	Aging	100
	Cancers	100
	Endocrine and metabolic diseases	100
	Endocrine system	100
	Cardiovascular diseases	100
	Signaling molecules and interaction	100
	Viral protein family	98.25

Table A 5 Number of orthologous IDs for each COG functional categories, after mapping the putative genes against the MetaHIT-2014 database



COG functional categories	Average number of orthologous IDs
unknown_funcat_id unknown	149910
[S] Function unknown	77302
[G] Carbohydrate transport and metabolism	58027
[R] General function prediction only	47641
[L] Replication, recombination and repair	41075
[M] Cell wall/membrane/envelope biogenesis	30818
[E] Amino acid transport and metabolism	30396
[K] Transcription	30403
[J] Translation, ribosomal structure and biogenesis	29237
[T] Signal transduction mechanisms	26121
[C] Energy production and conversion	26640
[V] Defense mechanisms	15434
[P] Inorganic ion transport and metabolism	14965
[O] Posttranslational modification, protein turnover, chaperones	14452
[H] Coenzyme transport and metabolism	11814
[F] Nucleotide transport and metabolism	10741
[I] Lipid transport and metabolism	8656
[U] Intracellular trafficking, secretion, and vesicular transport	8212
[N] Cell motility	6162
[D] Cell cycle control, cell division, chromosome partitioning	5955
[Q] Secondary metabolites biosynthesis, transport and catabolism	3949
[Z] Cytoskeleton	363
[B] Chromatin structure and dynamics	27
[W] Extracellular structures	22
[A] RNA processing and modification	10
[Y] Nuclear structure	0

Table A 6 Statistical results of PERMANOVA analysis for filters with FDR<0.3

FILTER	DB	R^2	P-VALUE	SIGNIF	FDR	SIGNIF(FDR)
COMPARISON: H-H						
-						
COMPARISON: H-IBD (CD+UC)						
F4	funcatKEGG.L3	2.17%	0.072	.	0.261	
F4	funcat	2.38%	0.086	.	0.284	
COMPARISON: H-CD						
F6	Species	5.21%	0.035	*	0.189	
F6	Family	5.35%	0.036	*	0.202	
F6	Class	7.81%	0.026	*	0.251	
F6	Genus	5.20%	0.05	*	0.28	
F9	funcatKEGG.L2.default	23.66%	0.001	***	0.014	*
F9	funcatKEGG.L4	21.06%	0.001	***	0.014	*
F9	orthidsEggNOG4.5	19.02%	0.001	***	0.014	*

F9	orthidsKEGG	20.83%	0.002	**	0.021	*
F9	orthids	19.82%	0.001	***	0.028	*
F9	funcatKEGG.L1	29.08%	0.001	***	0.029	*
F9	funcatKEGG.L3	22.97%	0.001	***	0.029	*
F9	funcat	26.06%	0.003	**	0.04	*
F9	funcatEggNOG4.5	25.59%	0.005	**	0.045	*
F9	Species	8.39%	0.042	*	0.189	
COMPARISON: H-UC						
F15	funcat	6.31%	0.065	.	0.244	
F17	orthidsEggNOG4.5	9.93%	0.006	**	0.032	*
F17	orthids	8.54%	0.022	*	0.123	
F17	orthidsKEGG	8.23%	0.033	*	0.154	
F17	funcatKEGG.L4	8.12%	0.043	*	0.174	
F17	funcat	8.23%	0.104		0.284	
F19	funcat	13.42%	0.033	*	0.165	
F19	funcatKEGG.L4	10.31%	0.048	*	0.174	
F19	funcatKEGG.L2.default	11.43%	0.047	*	0.181	
F19	funcatKEGG.L3	12.02%	0.044	*	0.199	
F19	orthidsKEGG	10.24%	0.057	.	0.2	
F19	funcatKEGG.L1	14.13%	0.062	.	0.257	
F19	orthids	8.57%	0.072	.	0.287	
F21	Genus	9.24%	0.018	*	0.168	
F21	Species	8.20%	0.032	*	0.189	
F21	Family	9.42%	0.031	*	0.202	
F21	Class	12.37%	0.018	*	0.251	
COMPARISON: CD-UC (INCL HEALTHY)						
F22	funcat	2.60%	0.096	.	0.284	
F23	orthidsKEGG	6.49%	0.003	**	0.021	*
F23	funcatEggNOG4.5	15.26%	0.001	***	0.027	*
F23	orthidsEggNOG4.5	5.80%	0.003	**	0.027	*
F23	funcatKEGG.L4	6.45%	0.003	**	0.029	*
F23	funcat	9.30%	0.004	**	0.04	*
F23	orthids	5.95%	0.007	**	0.049	*
F23	funcatKEGG.L2.default	8.60%	0.007	**	0.061	.
F23	funcatKEGG.L3	7.68%	0.007	**	0.068	.
F23	funcatKEGG.L1	9.66%	0.008	**	0.077	.
F25	funcatKEGG.L2.default	6.56%	0.012	*	0.065	.
F25	funcatKEGG.L1	6.60%	0.02	*	0.116	
F25	funcat	5.63%	0.028	*	0.165	
F25	funcatKEGG.L4	3.80%	0.048	*	0.174	
F25	orthidsKEGG	3.74%	0.045	*	0.18	
F25	Species	4.71%	0.022	*	0.189	
F25	funcatKEGG.L3	4.79%	0.033	*	0.191	
F25	Genus	5.02%	0.028	*	0.196	
F25	Family	4.96%	0.024	*	0.202	
F25	Class	6.81%	0.02	*	0.251	
F25	orthids	3.08%	0.082	.	0.287	

F26	orthidsKEGG	6.65%	0.003	**	0.021	*
F26	funcatKEGG.L4	6.65%	0.004	**	0.029	*
F26	orthidsEggNOG4.5	5.58%	0.005	**	0.032	*
F26	orthids	6.13%	0.005	**	0.047	*
F26	funcatKEGG.L2.default	8.05%	0.009	**	0.061	.
F26	funcat	10.34%	0.009	**	0.067	.
F26	funcatKEGG.L3	7.25%	0.011	*	0.08	.
F26	funcatKEGG.L1	9.01%	0.013	*	0.094	.
F26	funcatEggNOG4.5	8.18%	0.025	*	0.169	
COMPARISON: CD-UC						
F27	funcatKEGG.L2.default	22.20%	0.001	***	0.014	*
F27	funcatKEGG.L4	16.06%	0.001	***	0.014	*
F27	orthidsEggNOG4.5	13.63%	0.001	***	0.014	*
F27	orthidsKEGG	16.02%	0.001	***	0.021	*
F27	funcatEggNOG4.5	25.11%	0.002	**	0.027	*
F27	orthids	14.57%	0.002	**	0.028	*
F27	funcat	20.94%	0.002	**	0.04	*
F27	funcatKEGG.L1	25.10%	0.003	**	0.044	*
F27	funcatKEGG.L3	19.16%	0.003	**	0.044	*
F27	Species	11.72%	0.006	**	0.081	.
F27	Genus	11.11%	0.016	*	0.168	
F27	Family	11.08%	0.016	*	0.202	
F28	Species	12.37%	0.003	**	0.081	.
F28	Genus	12.00%	0.005	**	0.14	
F28	Family	12.62%	0.008	**	0.202	
COMPARISON: CD-CD (REM vs REL)						
-						
COMPARISON: UC-UC (REM vs REL)						
F32	funcatEggNOG4.5	28.60%	0.066	.	0.297	
F33	funcatEggNOG4.5	27.50%	0.046	*	0.248	
F34	funcatKEGG.L4	8.31%	0.013	*	0.075	.
F34	orthidsKEGG	8.25%	0.022	*	0.123	
F34	orthids	7.02%	0.028	*	0.131	
F34	funcatKEGG.L2.default	8.55%	0.047	*	0.181	
F34	funcatKEGG.L1	10.01%	0.038	*	0.184	
F34	funcatKEGG.L3	7.79%	0.048	*	0.199	
F34	orthidsEggNOG4.5	6.22%	0.054	.	0.243	
F34	funcat	8.53%	0.059	.	0.244	
COMPARISON: CD PREDICTION						
-						
COMPARISON: UC PREDICTION						
-						

Significance level code: * is <=0,05 (<=5%) ** is <=0,01 (<=1%) *** is <=0,001 (<=0.1%)

Table A 7 Significant enriched KEGG metabolic pathways between healthy relatives of CD and CD patients (filter F9; p-values<0.05).

Pathway	Definition	Orthology.count	Coverage	pvalue
map02040	Flagellar assembly	32 (up:3, down:29)	80.0%	2.03E-19
map02030	Bacterial chemotaxis	19 (up:4, down:15)	73.1%	8.18E-11
map00550	Peptidoglycan biosynthesis	26 (up:20, down:6)	53.1%	1.12E-09
map00680	Methane metabolism	60 (up:12, down:48)	32.3%	7.01E-09
map01200	Carbon metabolism	94 (up:43, down:51)	26.6%	3.40E-08
map01230	Biosynthesis of amino acids	67 (up:39, down:28)	28.9%	1.28E-07
map00620	Pyruvate metabolism	37 (up:14, down:23)	35.9%	2.67E-07
map00720	Carbon fixation pathways in prokaryotes	35 (up:17, down:18)	34.0%	2.48E-06
map02020	Two-component system	113 (up:57, down:56)	22.8%	8.23E-06
map00540	Lipopolysaccharide biosynthesis	22 (up:9, down:13)	37.3%	3.55E-05
map00640	Propanoate metabolism	32 (up:11, down:21)	29.6%	1.40E-04
map03030	DNA replication	21 (up:9, down:12)	35.0%	1.53E-04
map00020	Citrate cycle (TCA cycle)	20 (up:8, down:12)	34.5%	2.75E-04
map00670	One carbon pool by folate	14 (up:10, down:4)	38.9%	5.63E-04
map00500	Starch and sucrose metabolism	30 (up:18, down:12)	26.8%	1.42E-03
map00983	Drug metabolism - other enzymes	12 (up:8, down:4)	37.5%	2.00E-03
map00400	Phenylalanine, tyrosine and tryptophan biosynthesis	21 (up:8, down:13)	29.2%	2.31E-03
map00860	Porphyrin and chlorophyll metabolism	32 (up:21, down:11)	24.6%	4.30E-03
map03430	Mismatch repair	14 (up:9, down:5)	31.8%	5.09E-03
map00240	Pyrimidine metabolism	25 (up:15, down:10)	25.3%	7.70E-03
map02026	Biofilm formation - Escherichia coli	17 (up:12, down:5)	27.9%	9.46E-03
map00230	Purine metabolism	45 (up:30, down:15)	21.6%	1.09E-02
map00300	Lysine biosynthesis	13 (up:11, down:2)	28.9%	1.61E-02
map00970	Aminoacyl-tRNA biosynthesis	17 (up:16, down:1)	26.2%	1.80E-02
map00630	Glyoxylate and dicarboxylate metabolism	24 (up:10, down:14)	23.8%	1.88E-02
map00450	Selenocompound metabolism	10 (up:8, down:2)	31.3%	1.91E-02
map00473	D-Alanine metabolism	3 (up:2, down:1)	60.0%	2.90E-02
map01503	Cationic antimicrobial peptide (CAMP) resistance	14 (up:7, down:7)	25.9%	3.23E-02
map00633	Nitrotoluene degradation	7 (up:2, down:5)	33.3%	3.35E-02
map00250	Alanine, aspartate and glutamate metabolism	17 (up:14, down:3)	24.3%	3.57E-02
map04112	Cell cycle - Caulobacter	9 (up:7, down:2)	29.0%	4.06E-02
map00480	Glutathione metabolism	13 (up:13)	25.5%	4.34E-02
map03440	Homologous recombination	17 (up:11, down:6)	23.6%	4.56E-02

Coverage and Orthology.count refer to KEGG orthologous IDs mapped within the pathway.

Table A 8 Differentially expressed orthologous genes between healthy relatives of UC and UC patients classified in functional categories (filter F17; DEG genes at FDR<0.05)

	Up (log2)	Down (log2)
[A] RNA processing and modification	0	0
[B] Chromatin structure and dynamics	0	0

[C] Energy production and conversion	7.3	7.2
[D] Cell cycle control, cell division, chromosome partitioning	4.6	3.2
[E] Amino acid transport and metabolism	7.7	6.8
[F] Nucleotide transport and metabolism	6.6	5.7
[G] Carbohydrate transport and metabolism	7.7	7.2
[H] Coenzyme transport and metabolism	6.3	6.4
[I] Lipid transport and metabolism	5.4	5
[J] Translation, ribosomal structure and biogenesis	7.6	6.4
[K] Transcription	7.5	6.8
[L] Replication, recombination and repair	7.2	6.6
[M] Cell wall/membrane/envelope biogenesis	7	7.1
[N] Cell motility	0	2
[O] Posttranslational modification, protein turnover, chaperones	6.4	6.1
[P] Inorganic ion transport and metabolism	7	6.5
[Q] Secondary metabolites biosynthesis, transport and catabolism	3.9	3.7
[R] General function prediction only	0	0
[S] Function unknown	9.4	9.8
[T] Signal transduction mechanisms	6.2	5.5
[U] Intracellular trafficking, secretion, and vesicular transport	4.2	4.5
[V] Defense mechanisms	5.7	5.7
[W] Extracellular structures	0	0
[Y] Nuclear structure	0	0
[Z] Cytoskeleton	0	0
UNKNOWN	6.5	7.7

Table A 9 Significant enriched KEGG metabolic pathways between UC and CD patients (filter F27; p-values<0.05).

Pathway	Definition	Orthology.count	Coverage	pvalue
map02040*	Flagellar assembly	24 (up:3, down:21)	60%	8.68E-16
map02030*	Bacterial chemotaxis	16 (up:5, down:11)	62%	3.46E-11
map01230*	Biosynthesis of amino acids	45 (up:23, down:22)	19%	5.00E-07
map02020*	Two-component system	73 (up:38, down:35)	15%	1.35E-05
map00620*	Pyruvate metabolism	23 (up:12, down:11)	22%	3.12E-05
map00720*	Carbon fixation pathways in prokaryotes	23 (up:15, down:8)	22%	3.12E-05
map01200*	Carbon metabolism	55 (up:41, down:14)	16%	3.22E-05
map00240*	Pyrimidine metabolism	21 (up:16, down:5)	21%	1.49E-04
map00630*	Glyoxylate and dicarboxylate metabolism	21 (up:9, down:12)	21%	2.00E-04
map00020*	Citrate cycle (TCA cycle)	13 (up:9, down:4)	22%	1.54E-03
map00540*	Lipopolysaccharide biosynthesis	13 (up:8, down:5)	22%	1.82E-03
map00670*	One carbon pool by folate	9 (up:6, down:3)	25%	3.66E-03
map00400*	Phenylalanine, tyrosine and tryptophan biosynthesis	14 (up:2, down:12)	19%	4.23E-03
map00450*	Selenocompound metabolism	8 (up:4, down:4)	25%	6.04E-03
map00010	Glycolysis / Gluconeogenesis	17 (up:13, down:4)	17%	9.71E-03
map00500*	Starch and sucrose metabolism	18 (up:11, down:7)	16%	1.04E-02
map00230*	Purine metabolism	29 (up:21, down:8)	14%	1.10E-02

map00860*	Porphyrin and chlorophyll metabolism	20 (up:18, down:2)	15%	1.15E-02
map00710	Carbon fixation in photosynthetic organisms	8 (up:6, down:2)	22%	1.26E-02
map00640*	Propanoate metabolism	17 (up:6, down:11)	16%	1.53E-02
map02026*	Biofilm formation - Escherichia coli	11 (up:7, down:4)	18%	1.84E-02
map01210	2-Oxocarboxylic acid metabolism	13 (up:6, down:7)	16%	2.98E-02
map03070	Bacterial secretion system	12 (up:5, down:7)	16%	3.07E-02
map00480*	Glutathione metabolism	9 (up:9)	18%	3.55E-02
map03430*	Mismatch repair	8 (up:5, down:3)	18%	3.94E-02
map00300*	Lysine biosynthesis	8 (up:6, down:2)	18%	4.43E-02
map00770	Pantothenate and CoA biosynthesis	7 (up:5, down:2)	18%	4.91E-02

* pathways shared with HR-CD comparison set.

Chapter 9.

References

9 References

Abbas-Aghababazadeh, F., Li, Q. and Fridley, B. L. (2018) ‘Comparison of normalization approaches for gene expression studies completed with high-throughput sequencing’, *PloS one*, 13(10), p. e0206312. doi: 10.1371/journal.pone.0206312.

Abu-Ali, G. S. *et al.* (2018) ‘Metatranscriptome of human faecal microbial communities in a cohort of adult men’, *Nature Microbiology*. Springer US, 3(3), pp. 356–366. doi: 10.1038/s41564-017-0084-4.

Abubucker, S. *et al.* (2012) ‘HUMAnN - Metabolic reconstruction for metagenomic data and its application to the human microbiome’, *PLoS Computational Biology*, 8(6). doi: 10.1371/journal.pcbi.1002358.

Aguiar-pulido, V. *et al.* (2016) ‘Metagenomics, Metatranscriptomics, and Metabolomics Approaches for Microbiome Analysis’, 12, pp. 5–16. doi: 10.4137/EBO.S36436.TYPE.

Andersson, A. F. *et al.* (2008) ‘Comparative analysis of human gut microbiota by barcoded pyrosequencing’, *PLoS ONE*. doi: 10.1371/journal.pone.0002836.

Andrews, S. (2010) ‘FastQC A Quality Control tool for High Throughput Sequence Data’, *http://www.bioinformatics.babraham.ac.uk/projects/fastqc/*. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (Accessed: 25 November 2013).

Aronesty, E. (2013) *Comparison of Sequencing Utility Programs*, *The Open Bioinformatics*. doi: 10.2174/1875036201307010001.

Bashiardes, S., Zilberman-Schapira, G. and Elinav, E. (2016) ‘Use of metatranscriptomics in microbiome research’, *Bioinformatics and Biology Insights*, 10, pp. 19–25. doi: 10.4137/BBI.S34610.

Bernalier, A. *et al.* (1996) ‘Ruminococcus hydrogenotrophicus sp. nov., a new H2CO2-utilizing acetogenic bacterium isolated from human feces’, *Archives of Microbiology*, 166(3), pp. 176–183. doi: 10.1007/s002030050373.

Best, W. R. *et al.* (1976) ‘Development of a Crohn’s disease activity index.

National Cooperative Crohn’s Disease Study.’, *Gastroenterology*, 70(3), pp. 439–44. doi: 10.1016/S0016-5085(76)80163-1.

Bokulich, N. A. *et al.* (2013) ‘Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing.’, *Genomics*. NIH Public Access, 10(1), pp. 151–170. doi: 10.1146/annurev-genom-090711-163814.The.

Breitbart, M. *et al.* (2008) ‘Viral diversity and dynamics in an infant gut’, *Research in Microbiology*. doi: 10.1016/j.resmic.2008.04.006.

Brown, K. *et al.* (2012) ‘Diet-induced dysbiosis of the intestinal microbiota and the effects on immunity and disease’, *Nutrients*. doi: 10.3390/nu4081095.

Buchfink, B., Xie, C. and Huson, D. H. (2014) ‘Fast and sensitive protein alignment using DIAMOND’, *Nature methods*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 12(1), pp. 59–60. doi: 10.1038/nmeth.3176.

Burge, S. W. *et al.* (2013) ‘Rfam 11.0: 10 years of RNA families’, *Nucleic acids research*, 41(Database issue), pp. D226-32. doi: 10.1093/nar/gks1005.

Caporaso, J. G. *et al.* (2010) ‘QIIME allows analysis of high-throughput community sequencing data.’, *Nature methods*, 7(5), pp. 335–336. doi: 10.1038/nmeth.f.303.

Caporaso, J. G. *et al.* (2011) ‘Moving pictures of the human microbiome’, *Genome Biology*. doi: 10.1186/gb-2011-12-5-r50.

Cardona, S. *et al.* (2012) ‘Storage conditions of intestinal microbiota matter in metagenomic analysis.’, *BMC microbiology*, 12(1), p. 158. doi: 10.1186/1471-2180-12-158.

Caspi, R. *et al.* (2018) ‘The MetaCyc database of metabolic pathways and enzymes’, *Nucleic Acids Research*. Oxford University Press, 46(D1), pp. D633–D639. doi: 10.1093/nar/gkx935.

Chan, P. P. and Lowe, T. M. (2009) ‘GtRNAdb: a database of transfer RNA genes detected in genomic sequence.’, *Nucleic acids research*, 37(Database issue), pp. D93-7. doi: 10.1093/nar/gkn787.

Chassaing, B., Etienne-Mesmin, L. and Gewirtz, A. T. (2014) ‘Microbiota-liver axis in hepatic disease’, *Hepatology*. doi: 10.1002/hep.26494.

Le Chatelier, E. *et al.* (2013) ‘Richness of human gut microbiome correlates with metabolic markers.’, *Nature*, 500(7464), pp. 541–6. doi: 10.1038/nature12506.

Claesson, M. J. *et al.* (2012) ‘Gut microbiota composition correlates with diet and health in the elderly’, *Nature*, 488(7410), pp. 178–184. doi: 10.1038/nature11319.

Collins, F. S. *et al.* (2004) ‘Finishing the euchromatic sequence of the human genome’, *Nature*. doi: 10.1038/nature03001.

Cotillard, A. *et al.* (2013) ‘Dietary intervention impact on gut microbial gene richness.’, *Nature*, 500(7464), pp. 585–8. doi: 10.1038/nature12480.

Cryan, J. F. and Dinan, T. G. (2012) ‘Mind-altering microorganisms: The impact of the gut microbiota on brain and behaviour’, *Nature Reviews Neuroscience*. doi: 10.1038/nrn3346.

Darzi, Y. *et al.* (2018) ‘IPath3.0: Interactive pathways explorer v3’, *Nucleic Acids Research*. Oxford University Press, 46(W1), pp. W510–W513. doi: 10.1093/nar/gky299.

Davis, M. P. A. A. *et al.* (2013) ‘Kraken: a set of tools for quality control and analysis of high-throughput sequence data.’, *Methods (San Diego, Calif.)*, 63(1), pp. 41–9. doi: 10.1016/j.ymeth.2013.06.027.

DeSantis, T. Z. *et al.* (2006) ‘Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB’, *Applied and Environmental Microbiology*, 72(7), pp. 5069–5072. doi: 10.1128/AEM.03006-05.

Dominguez-Bello, M. G. *et al.* (2010) ‘Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns’, *Proceedings of the National Academy of Sciences*, 107(26), pp. 11971–11975. doi: 10.1073/pnas.1002601107.

Duchmann, R. *et al.* (1995) ‘Tolerance exists towards resident intestinal flora but is broken in active inflammatory bowel disease (IBD)’, *Clinical and experimental immunology*. doi: 10.1111/j.1365-2249.1995.tb03836.x.

Edgar, R. C. (2010) ‘Search and clustering orders of magnitude faster than BLAST.’, *Bioinformatics (Oxford, England)*, 26(19), pp. 2460–1. doi: 10.1093/bioinformatics/btq461.

Fakhoury, M. *et al.* (2014) ‘Inflammatory bowel disease: Clinical aspects and treatments’, *Journal of Inflammation Research*. Dove Press, 7(1), pp. 113–120. doi: 10.2147/JIR.S65979.

Feng, Y. *et al.* (2018) ‘Short-Chain Fatty Acids Manifest Stimulative and Protective Effects on Intestinal Barrier Function Through the Inhibition of NLRP3 Inflammasome and Autophagy’, *Cellular Physiology and Biochemistry*. doi: 10.1159/000492853.

De Filippo, C. *et al.* (2010) ‘Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa.’, *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.1005963107.

Foxman, B. and Martin, E. T. (2015) ‘Practice of Epidemiology Use of the Microbiome in the Practice of Epidemiology : A Primer on -Omic Technologies’, *American Journal of Epidemiology*. doi: 10.1093/aje/kwv102.

Frank, D. N. *et al.* (2007) ‘Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases.’, *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.0706625104.

Frank, D. N. and Pace, N. R. (2008) ‘Gastrointestinal microbiology enters the metagenomics era’, *Current Opinion in Gastroenterology*. doi: 10.1097/MOG.0b013e3282f2b0e8.

Franzosa, E. A. *et al.* (2018) ‘Functionally profiling metagenomes and metatranscriptomes at species-level resolution’, *Nature methods*. Springer US, accepted(November). doi: 10.1038/s41592-018-0176-y.

Franzosa, E. a *et al.* (2014) ‘Relating the metatranscriptome and metagenome of the human gut.’, *Proceedings of the National Academy of Sciences of the United States of America*, 111(22), pp. E2329-38. doi: 10.1073/pnas.1319284111.

Frazee, A. C. *et al.* (2015) 'Polyester: simulating RNA-seq datasets with differential transcript expression', *Bioinformatics*. , 31(17), pp. 2778–2784. doi: 10.1093/bioinformatics/btv272.

Fu, L. *et al.* (2012) 'CD-HIT: accelerated for clustering the next-generation sequencing data.', *Bioinformatics (Oxford, England)*, 28(23), pp. 3150–3152. doi: 10.1093/bioinformatics/bts565.

Garrett, W. S., Gordon, J. I. and Glimcher, L. H. (2010) 'Homeostasis and Inflammation in the Intestine', *Cell*. doi: 10.1016/j.cell.2010.01.023.

Gensollen, T. *et al.* (2016) 'How colonization by microbiota in early life shapes the immune 1. Gensollen T, Iyer SS, Kasper DL, Blumberg RS, Medical H. How colonization by microbiota in early life shapes the immune system. Science. 2016;352(6285):539-544. doi:10.1126/science.aad9378.', *Science (New York, N.Y.)*. NIH Public Access, 352(6285), pp. 539–544. doi: 10.1126/science.aad9378.How.

Giannoukos, G. *et al.* (2012) 'Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes.', *Genome Biology*. BioMed Central Ltd, 13(3), p. R23. doi: 10.1186/gb-2012-13-3-r23.

Van Der Giessen, J. *et al.* (2019) 'Modulation of cytokine patterns and microbiome during pregnancy in IBD', *Gut*. doi: 10.1136/gutjnl-2019-318263.

González, E. and Joly, S. (2013) 'Impact of RNA-seq attributes on false positive rates in differential expression analysis of de novo assembled transcriptomes.', *BMC research notes*, 6(1), p. 503. doi: 10.1186/1756-0500-6-503.

Gosalbes, M. J. *et al.* (2011) 'Metatranscriptomic approach to analyze the functional human gut microbiota.', *PLoS ONE*. Edited by L. Quintana-Murci. Public Library of Science, 6(3), p. e17447. doi: 10.1371/journal.pone.0017447.

Grice, E. A. and Segre, J. A. (2012) 'The Human Microbiome: Our Second Genome', *Annual Review of Genomics and Human Genetics*. NIH Public Access, 13(1), pp. 151–170. doi: 10.1146/annurev-genom-090711-163814.

Haas, B. J. *et al.* (2013) 'De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis.', *Nature* *protocols*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 8(8), pp. 1494–1512. doi: 10.1038/nprot.2013.084.

Hall, A. B. *et al.* (2017) ‘A novel Ruminococcus gnavus clade enriched in inflammatory bowel disease patients’, *Genome Medicine*. Genome Medicine, 9(1), pp. 1–12. doi: 10.1186/s13073-017-0490-5.

He, Q. *et al.* (2017) ‘Two distinct metacommunities characterize the gut microbiota in Crohn’s disease patients’, *GigaScience*. doi: 10.1093/gigascience/gix050.

Henke, M. T. *et al.* (2019) ‘Ruminococcus gnavus, a member of the human gut microbiome associated with Crohn’s disease, produces an inflammatory polysaccharide’, *Proceedings of the National Academy of Sciences of the United States of America*, 116(26), pp. 12672–12677. doi: 10.1073/pnas.1904099116.

Hernando-Harder, A. C. *et al.* (2010) ‘Colonic responses to gas loads in subgroups of patients with abdominal bloating’, *American Journal of Gastroenterology*. doi: 10.1038/ajg.2010.75.

Hill, D. A. and Artis, D. (2010) ‘Intestinal Bacteria and the Regulation of Immune Cell Homeostasis’, *Annual Review of Immunology*. doi: 10.1146/annurev-immunol-030409-101330.

Hooper, L. V. and MacPherson, A. J. (2010) ‘Immune adaptations that maintain homeostasis with the intestinal microbiota’, *Nature Reviews Immunology*. doi: 10.1038/nri2710.

Huang, H. *et al.* (2014) ‘Multi-omics analysis of inflammatory bowel disease (IBD)’, *Immunology Letters*. Elsevier B.V., 162(2), pp. 62–68. doi: 10.1016/j.imlet.2014.07.014.

Huang, W. *et al.* (2012) ‘ART: a next-generation sequencing read simulator’, *Bioinformatics*. , 28(4), pp. 593–594. doi: 10.1093/bioinformatics/btr708.

Huerta-Cepas, J. *et al.* (2016) ‘EGGNOG 4.5: A hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences’, *Nucleic Acids Research*, 44(D1), pp. D286–D293. doi: 10.1093/nar/gkv1248.

Huttenhower, C. *et al.* (2012) ‘Structure, function and diversity of the healthy human microbiome.’, *Nature*. Nature Publishing Group, 486(7402), pp. 207–14. doi: 10.1038/nature11234.

Huttenhower, C. *et al.* (2015) ‘MetaPhlAn2 for enhanced metagenomic taxonomic profiling’, *Nature Methods*, 12(10), pp. 902–903. doi: 10.1038/nmeth.3589.

Illumina (2012) ‘An Introduction to Next-Generation Sequencing Technology’, pp. 1–12. doi: Pub No. 770-2012-008.

Imhann, F. *et al.* (2018) ‘Interplay of host genetics and gut microbiota underlying the onset and clinical presentation of inflammatory bowel disease’, *Gut*, 67(1), pp. 108–119. doi: 10.1136/gutjnl-2016-312135.

Jakobsson, H. E. *et al.* (2014) ‘Decreased gut microbiota diversity, delayed Bacteroidetes colonisation and reduced Th1 responses in infants delivered by Caesarean section’, *Gut*, 63(4), pp. 559–566. doi: 10.1136/gutjnl-2012-303249.

Jaszczyszyn, Y. *et al.* (2014) ‘Ten years of next-generation sequencing technology’, pp. 1–9. doi: 10.1016/j.tig.2014.07.001.

Jeraldo, P. *et al.* (2014) ‘IM-TORNADO: A tool for comparison of 16S reads from paired-end libraries’, *PLoS ONE*. Edited by P. J. Janssen, 9(12), p. e114804. doi: 10.1371/journal.pone.0114804.

KA., Wetterstrand, . (2018) *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)*. Available at: https://www.genome.gov/sequencingcosts/ (Accessed: 25 April 2018).

Kanehisa, M. *et al.* (2017) ‘KEGG: New perspectives on genomes, pathways, diseases and drugs’, *Nucleic Acids Research*, 45(D1), pp. D353--D361. doi: 10.1093/nar/gkw1092.

Kanehisa, M. *et al.* (2019) ‘New approach for understanding genome variations in KEGG’, *Nucleic Acids Research*. doi: 10.1093/nar/gky962.

Kanehisa, M. and Goto, S. (2000) ‘KEGG: kyoto encyclopedia of genes and genomes.’, *Nucleic acids research*, 28(1), pp. 27–30. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102409&tool=pmcent

rez&rendertype=abstract (Accessed: 10 July 2014).

Kaplan, G. G. (2015) ‘The global burden of IBD: From 2015 to 2025’, *Nature Reviews Gastroenterology and Hepatology*. Nature Publishing Group, 12(12), pp. 720–727. doi: 10.1038/nrgastro.2015.150.

Kaser, A., Zeissig, S. and Blumberg, R. S. (2010) ‘Inflammatory Bowel Disease’, *Annual Review of Immunology*. doi: 10.1146/annurev-immunol-030409-101225.

Kchouk, M., Gibrat, J. F. and Elloumi, M. (2017) ‘Generations of Sequencing Technologies: From First to Next Generation’, *Biology and Medicine*. OMICS International, 09(03). doi: 10.4172/0974-8369.1000395.

Kim, J. *et al.* (2016) ‘FMAP: Functional Mapping and Analysis Pipeline for metagenomics and metatranscriptomics studies’, *BMC Bioinformatics*. BMC Bioinformatics, 17(1), p. 420. doi: 10.1186/s12859-016-1278-0.

Klingenberg, H. and Meinicke, P. (2017) ‘How to normalize metatranscriptomic count data for differential expression analysis’, *PeerJ*, 5, p. e3859. doi: 10.7717/peerj.3859.

Knight, R. *et al.* (2012) ‘Unlocking the potential of metagenomics through replicated experimental design’, *Nature Biotechnology*. doi: 10.1038/nbt.2235.

Knight, R. *et al.* (2018) ‘Best practices for analysing microbiomes’, *Nature Reviews Microbiology*. Springer US, 16(July), pp. 1–13. doi: 10.1038/s41579-018-0029-9.

Koeth, R. A. *et al.* (2013) ‘Intestinal microbiota metabolism of l-carnitine, a nutrient in red meat, promotes atherosclerosis’, *Nature Medicine*. doi: 10.1038/nm.3145.

Kopylova, E., Noé, L. and Touzet, H. (2012) ‘SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data.’, *Bioinformatics (Oxford, England)*, 28(24), pp. 3211–7. doi: 10.1093/bioinformatics/bts611.

Kostic, A. D., Xavier, R. J. and Gevers, D. (2014) ‘The microbiome in inflammatory bowel disease: Current status and the future ahead’, *Gastroenterology*. doi: 10.1053/j.gastro.2014.02.009.

Kultima, J. R. *et al.* (2016) ‘MOCAT2: A metagenomic assembly, annotation and profiling framework’, *Bioinformatics*. Oxford University Press, 32(16), pp. 2520–2523. doi: 10.1093/bioinformatics/btw183.

Kvalseth, T. O. (2015) ‘Evenness indices once again: critical analysis of properties’, *SpringerPlus*. ???, 4(232), p. 12. doi: 10.1186/s40064-015-0944-4.

Lahti, L. and Shetty, S. (2012) ‘microbiome R package’.

Lederberg, B. J. and McCray, A. T. (2001) ‘’ Ome Sweet ’ Omics-- A Genealogical Treasury of Words’, *The Scientist*, 15(7), p. 8. Available at: https://lhncbc.nlm.nih.gov/system/files/pub2001047.pdf.

Lee, S. J. and Maizels, R. M. (2014) ‘Inflammatory Bowel Disease’, *Evolution, Medicine, and Public Health*, 2014(1), p. 95. doi: 10.1093/emph/eou017.

Lee, Z. M. P. *et al.* (2009) ‘rrnDB: documenting the number of rRNA and tRNA genes in bacteria and archaea’, *Nucleic Acids Res*, 37(SUPPL. 1), pp. D489-493. doi: 10.1093/nar/gkn689.

Leimena, M. M. *et al.* (2013) ‘A comprehensive metatranscriptome analysis pipeline and its validation using human small intestine microbiota datasets.’, *BMC genomics*, 14(1), p. 530. doi: 10.1186/1471-2164-14-530.

Letunic, I. *et al.* (2008) ‘iPath: interactive exploration of biochemical pathways and networks’, *Trends in biochemical sciences*, pp. 101–103. Available at: http://libgen.org/scimag3/10.1016/j.tibs.2008.01.001.pdf (Accessed: 7 October 2014).

Ley, R. E., Peterson, D. A. and Gordon, J. I. (2006) ‘Ecological and evolutionary forces shaping microbial diversity in the human intestine’, *Cell*. Elsevier, 124(4), pp. 837–848. doi: 10.1016/j.cell.2006.02.017.

Li, H. (2013) ‘Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM’, *arXiv preprint arXiv*, 00(00), p. 3. doi: arXiv:1303.3997 [q-bio.GN].

Li, J. *et al.* (2014) ‘An integrated catalog of reference genes in the human gut microbiome’, *Nature Biotechnology*, 32(8), pp. 834–841. doi: 10.1038/nbt.2942.

Li, M. *et al.* (2018) ‘The anti-inflammatory effects of short chain fatty acids on lipopolysaccharide- or tumor necrosis factor α-stimulated endothelial cells via activation of GPR41/43 and inhibition of HDACs', *Frontiers in Pharmacology*, 9(MAY), pp. 1–12. doi: 10.3389/fphar.2018.00533.

Li, R. *et al.* (2009) 'SOAP2: an improved ultrafast tool for short read alignment', *Bioinformatics*. , 25(15), pp. 1966–1967. doi: 10.1093/bioinformatics/btp336.

Liu, S., Qin, P. and Wang, J. (2019) 'High-Fat Diet Alters the Intestinal Microbiota in Streptozotocin-Induced Type 2 Diabetic Mice', *Microorganisms*, 7(6), p. 176. doi: 10.3390/microorganisms7060176.

Liu, T. *et al.* (2012) 'Short-Chain fatty acids suppress lipopolysaccharide-Induced production of nitric oxide and proinflammatory cytokines through inhibition of NF-?B Pathway in RAW264.7 cells', *Inflammation*. doi: 10.1007/s10753-012-9484-z.

Lloyd-Price, J. *et al.* (2019) 'Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases', *Nature*, 569(7758), pp. 655–662. doi: 10.1038/s41586-019-1237-9.

Lloyd-Price, J., Abu-Ali, G. and Huttenhower, C. (2016) 'The healthy human microbiome', *Genome Medicine*. BioMed Central, p. 51. doi: 10.1186/s13073-016-0307-y.

Love, M. I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biology*, 15(12), pp. 1–21. doi: 10.1186/s13059-014-0550-8.

Macpherson, A. *et al.* (1996) 'Mucosal antibodies in inflammatory bowel disease are directed against intestinal bacteria', *Gut*. doi: 10.1136/gut.38.3.365.

Manichanh, C. *et al.* (2006) 'Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach', *Gut*. doi: 10.1136/gut.2005.073817.

Manichanh, C. *et al.* (2012) 'The gut microbiota in IBD', *Nature Reviews Gastroenterology and Hepatology*, 9(10), pp. 599–608. doi: 10.1038/nrgastro.2012.152.

Manichanh, C. *et al.* (2013) 'Anal gas evacuation and colonic microbiota in patients with flatulence: effect of diet', *Gut*, 63(3), pp. 401–408. doi: 10.1136/gutjnl-2012-303013.

Marchesi, J. R. and Ravel, J. (2015) 'The vocabulary of microbiome research: a proposal', *Microbiome*. BioMed Central, 3(1), p. 31. doi: 10.1186/s40168-015-0094-5.

Mardis, E. R. (2017) 'DNA sequencing technologies: 2006-2016', *Nature Protocols*, 12(2), pp. 213–218. doi: 10.1038/nprot.2016.182.

Martinez, X. *et al.* (2016) 'MetaTrans: an open-source pipeline for metatranscriptomics', *Scientific Reports*. Nature Publishing Group, 6, p. 26447. doi: 10.1038/srep26447.

McDonald, D. *et al.* (2012) 'An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea.', *The ISME journal*. Nature Publishing Group, 6(3), pp. 610–8. doi: 10.1038/ismej.2011.139.

McNulty, N. P. *et al.* (2011) 'The Impact of a Consortium of Fermented Milk Strains on the Gut Microbiome of Gnotobiotic Mice and Monozygotic Twins', *Science Translational Medicine*, 3(106), pp. 106ra106--106ra106. doi: 10.1126/scitranslmed.3002701.

Meehan, C. J. and Beiko, R. G. (2014) 'A phylogenomic view of ecological specialization in the lachnospiraceae, a family of digestive tract-associated bacteria', *Genome Biology and Evolution*. , 6(3), pp. 703–713. doi: 10.1093/gbe/evu050.

Mehta, R. S. *et al.* (2018) 'Stability of the human faecal microbiome in a cohort of adult men', *Nature Microbiology*, 3(3), pp. 347–355. doi: 10.1038/s41564-017-0096-0.

Meyer, F. *et al.* (2008) 'The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes.', *BMC bioinformatics*, 9, p. 386. doi: 10.1186/1471-2105-9-386.

Meyer, F. *et al.* (2012) 'Functional predictions from inference and observation in sequence-based inflammatory bowel disease research', *Genome Biology*, 13(9).

doi: 10.1186/gb-2012-13-9-169.

Morgan, X. C. *et al.* (2012) ‘Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment’, *Genome Biology*. BioMed Central, 13(9), p. R79. doi: 10.1186/gb-2012-13-9-r79.

Morgan, X. C. and Huttenhower, C. (2014) ‘Meta’omic analytic techniques for studying the intestinal microbiome’, *Gastroenterology*. Elsevier, Inc, 146(6), pp. 1437-1448.e1. doi: 10.1053/j.gastro.2014.01.049.

National Academies of Sciences, Engineering, and Medicine, . (2018) *Environmental Chemicals, the Human Microbiome, and Health Risk: A Research Strategy.* Washington, DC: The National Academies Press. doi: 10.17226/24960.

Ng, S. C. *et al.* (2017) ‘Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies’, *The Lancet*. Elsevier Ltd, 390(10114), pp. 2769–2778. doi: 10.1016/S0140-6736(17)32448-0.

Oksanen, J. (2015) ‘Vegan : ecological diversity’, p. 12. Available at: http://cran.r-project.org/web/packages/vegan/vignettes/diversity-vegan.pdf.

Oksanen, J. *et al.* (2018) ‘vegan: Community Ecology Package’. Available at: https://cran.r-project.org/package=vegan%7D.

Ottman, N. *et al.* (2012) ‘The function of our microbiota: who is out there and what do they do?’, *Frontiers in Cellular and Infection Microbiology*. Frontiers, 2, p. 104. doi: 10.3389/fcimb.2012.00104.

Pascal, V. *et al.* (2017) ‘A microbial signature for Crohn ’ s disease’, *Gut*, pp. 1–10. doi: 10.1136/gutjnl-2016-313235.

Patil, I. and Powell, C. (2018) ‘ggstatsplot: “ggplot2” Based Plots with Statistical Details’. doi: 10.5281/zenodo.2074621.

Peloquin, J. M. *et al.* (2016) ‘Characterization of candidate genes in inflammatory bowel disease–associated risk loci’, *JCI Insight*. American Society for Clinical Investigation, 1(13), p. e87899. doi: 10.1172/jci.insight.87899.

Peterson, J. *et al.* (2009) ‘The NIH Human Microbiome Project’, *Genome*

Research, 19(12), pp. 2317–2323. doi: 10.1101/gr.096651.109.

Png, C. W. *et al.* (2010) ‘Mucolytic bacteria with increased prevalence in IBD mucosa augment in vitro utilization of mucin by other bacteria’, *American Journal of Gastroenterology*. doi: 10.1038/ajg.2010.281.

Powell, S. *et al.* (2012) ‘eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges.’, *Nucleic acids research*, 40(Database issue), pp. D284-9. doi: 10.1093/nar/gkr1060.

Prescott, S. L. (2017) ‘History of medicine: Origin of the term microbiome and why it matters’, *Human Microbiome Journal*. Elsevier, 4, pp. 24–25. doi: 10.1016/j.humic.2017.05.004.

Lo Presti, A. *et al.* (2019) ‘Fecal and Mucosal Microbiota Profiling in Irritable Bowel Syndrome and Inflammatory Bowel Disease’, *Frontiers in Microbiology*, 10(July), pp. 1–14. doi: 10.3389/fmicb.2019.01655.

Pruesse, E. *et al.* (2007) ‘SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB.’, *Nucleic Acids Research*, 35(21), pp. 7188–7196. doi: 10.1093/nar/gkm864.

Qin, J. *et al.* (2010) ‘A human gut microbial gene catalogue established by metagenomic sequencing.’, *Nature*. Macmillan Publishers Limited. All rights reserved, 464(7285), pp. 59–65. doi: 10.1038/nature08821.

Qin, J. *et al.* (2012) ‘A metagenome-wide association study of gut microbiota in type 2 diabetes’, *Nature*, 490(7418), pp. 55–60. doi: 10.1038/nature11450.

R Core Team (2018) ‘R: A Language and envionment for statistical computing’. Available at: https://cran.r-project.org/.

Regner, E. H. *et al.* (2018) ‘Functional intraepithelial lymphocyte changes in inflammatory bowel disease and spondyloarthritis have disease specific correlations with intestinal microbiota’, *Arthritis Research and Therapy*. doi: 10.1186/s13075-018-1639-3.

Rho, M., Tang, H. and Ye, Y. (2010) ‘FragGeneScan: predicting genes in short and error-prone reads.’, *Nucleic Acids Research*. , 38(20), p. e191. doi:

10.1093/nar/gkq747.

Rhoads, A. and Au, K. F. (2015) ‘PacBio Sequencing and Its Applications’, *Genomics, Proteomics and Bioinformatics*. Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China, 13(5), pp. 278–289. doi: 10.1016/j.gpb.2015.08.002.

Ricotta, C. (2017) ‘Of beta diversity, variance, evenness, and dissimilarity’, *Ecology and Evolution*, 7(13), pp. 4835–4843. doi: 10.1002/ece3.2980.

Rinninella, E. *et al.* (2019) ‘What is the Healthy Gut Microbiota Composition? A Changing Ecosystem across Age, Environment, Diet, and Diseases’, *Microorganisms*, 7(1), p. 14. doi: 10.3390/microorganisms7010014.

Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2009) ‘edgeR: A Bioconductor package for differential expression analysis of digital gene expression data’, *Bioinformatics*, 26(1), pp. 139–140. doi: 10.1093/bioinformatics/btp616.

Sarrabayrouse, G. *et al.* (2015) ‘Microbiota-specific CD4CD8αα Tregs: Role in intestinal immune homeostasis and implications for IBD’, *Frontiers in Immunology*. doi: 10.3389/fimmu.2015.00522.

Scheppach, W. (1994) ‘Effects of short chain fatty acids on gut morphology and function’, in *Gut*. doi: 10.1136/gut.35.1_suppl.s35.

Schirmer, M. *et al.* (2019) ‘Microbial genes and pathways in inflammatory bowel disease’, *Nature Reviews Microbiology*, (CD). doi: 10.1038/s41579-019-0213-6.

Schloss, J. A. (2008) ‘How to get genomes at one ten-thousandth the cost’, *Nature Biotechnology*. doi: 10.1038/nbt1008-1113.

Schurch, N. J. *et al.* (2016) ‘How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?’, *RNA*, 22(6), pp. 839–851. doi: 10.1261/rna.053959.115.

Segata, N. *et al.* (2011) ‘Metagenomic biomarker discovery and explanation’, *Genome Biology*. doi: 10.1186/gb-2011-12-6-r60.

Selber-Hnativ, S. *et al.* (2017) ‘Human gut microbiota: Toward an ecology of disease', *Frontiers in Microbiology*, 8(JUL). doi: 10.3389/fmicb.2017.01265.

Sender, R., Fuchs, S. and Milo, R. (2016) 'Revised Estimates for the Number of Human and Bacteria Cells in the Body', *PLoS Biology*. Public Library of Science, 14(8), p. e1002533. doi: 10.1371/journal.pbio.1002533.

Serra, J. *et al.* (2002) 'Lipid-induced intestinal gas retention in irritable bowel syndrome', *Gastroenterology*. doi: 10.1053/gast.2002.35394.

Shi, Y. *et al.* (2011) 'Integrated metatranscriptomic and metagenomic analyses of stratified microbial assemblages in the open ocean.', *The ISME journal*. Nature Publishing Group, 5(6), pp. 999–1013. doi: 10.1038/ismej.2010.189.

Singh, W. (2008) 'Robustness of three hierarchical agglomerative clustering techniques for ecological data.', *October*, (October), p. 100. Available at: http://www.unuftp.is/static/files/rannsoknarritegrdir/WarshaSingh_MastersThesis.pdf (Accessed: 15 November 2017).

Sokol, H. *et al.* (2008) 'Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients', *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.0804812105.

Sokol, H. *et al.* (2013) 'Card9 mediates intestinal epithelial cell restitution, t-helper 17 responses, and control of bacterial infection in mice', *Gastroenterology*. doi: 10.1053/j.gastro.2013.05.047.

Stearns, J. C. *et al.* (2011) 'Bacterial biogeography of the human digestive tract', *Scientific Reports*, 1(1), p. 170. doi: 10.1038/srep00170.

Suzek, B. E. *et al.* (2015) 'UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches', *Bioinformatics*. doi: 10.1093/bioinformatics/btu739.

Suzuki, R. and Shimodaira, H. (2015) 'pvclust: Hierarchical Clustering with P-Values via Multiscale Bootstrap Resampling'. Available at: http://cran.r-project.org/package=pvclust.

Tap, J. *et al.* (2015) 'Gut microbiota richness promotes its stability upon increased dietary fibre intake in healthy adults', *Environmental Microbiology*, 17(12), pp. 4954–4964. doi: 10.1111/1462-2920.13006.

Tatusov, R. L. *et al.* (2001) 'The COG database : new developments in phylogenetic classification of proteins from complete genomes', 29(1), pp. 22–28. Available at: http://nar.oxfordjournals.org/content/29/1/22.full.pdf.

Tatusov, R. L., Koonin, E. V. and Lipman, D. J. (1997) 'A genomic perspective on protein families', *Science*. doi: 10.1126/science.278.5338.631.

The jamovi Project (2019) 'Jamovi'. Available at: https://www.jamovi.org.

Turnbaugh, P. J. *et al.* (2006) 'An obesity-associated gut microbiome with increased capacity for energy harvest', *Nature*. doi: 10.1038/nature05414.

Turnbaugh, P. J. *et al.* (2009) 'A core gut microbiome in obese and lean twins', *Nature*. doi: 10.1038/nature07540.

Tyakht, A. V. *et al.* (2013) 'Human gut microbiota community structures in urban and rural populations in Russia', *Nature Communications*. doi: 10.1038/ncomms3469.

Uniprot Consortium, T. (2019) 'UniProt: a worldwide hub of protein knowledge', *Nucleic acids research*. Oxford University Press, 47(D1), pp. D506–D515. doi: 10.1093/nar/gky1049.

Ursell, L. K. *et al.* (2012) 'Defining the human microbiome', *Nutrition Reviews*. NIH Public Access, 70(SUPPL. 1), pp. S38--44. doi: 10.1111/j.1753-4887.2012.00493.x.

Venegas, D. P. *et al.* (2019) 'Short chain fatty acids (SCFAs)mediated gut epithelial and immune regulation and its relevance for inflammatory bowel diseases', *Frontiers in Immunology*, 10(MAR). doi: 10.3389/fimmu.2019.00277.

Vinolo, M. A. R. *et al.* (2011) 'Regulation of inflammation by short chain fatty acids', *Nutrients*. doi: 10.3390/nu3100858.

Wagner, G. P., Kin, K. and Lynch, V. J. (2012) 'Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples', *Theory in Biosciences*, 131(4), pp. 281–285. doi: 10.1007/s12064-012-

0162-3.

Walters, W. A., Xu, Z. and Knight, R. (2014) ‘Meta-analyses of human gut microbes associated with obesity and IBD’, *FEBS Letters*. doi: 10.1016/j.febslet.2014.09.039.

Wheeler, D. A. *et al.* (2008) ‘The complete genome of an individual by massively parallel DNA sequencing’, *Nature*. doi: 10.1038/nature06884.

Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. Available at: http://ggplot2.org.

Wilke, A. *et al.* (2012) ‘The M5nr: a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools.’, *BMC bioinformatics*. BioMed Central, 13, p. 141. doi: 10.1186/1471-2105-13-141.

Wilke, A. *et al.* (2015) ‘A RESTful API for accessing microbial community data for MG-RAST’, *P Lo S Comput Biol*. Edited by P. P. Gardner. , 11(1), p. e1004008. doi: 10.1371/journal.pcbi.1004008.

Wood, D. E. and Salzberg, S. L. (2014) ‘Kraken : ultrafast metagenomic sequence classification using exact alignments’, *Genome biology*. Available at: http://genomebiology.com/content/pdf/gb-2014-15-3-r46.pdf.

Yamada, T. *et al.* (2011) ‘iPath2.0: interactive pathway explorer.’, *Nucleic Acids Research*, 39(Web Server issue), pp. W412–W415. doi: 10.1093/nar/gkr313.

Yatsunenko, T. *et al.* (2012) ‘Human gut microbiome viewed across age and geography’, *Nature*. doi: 10.1038/nature11053.

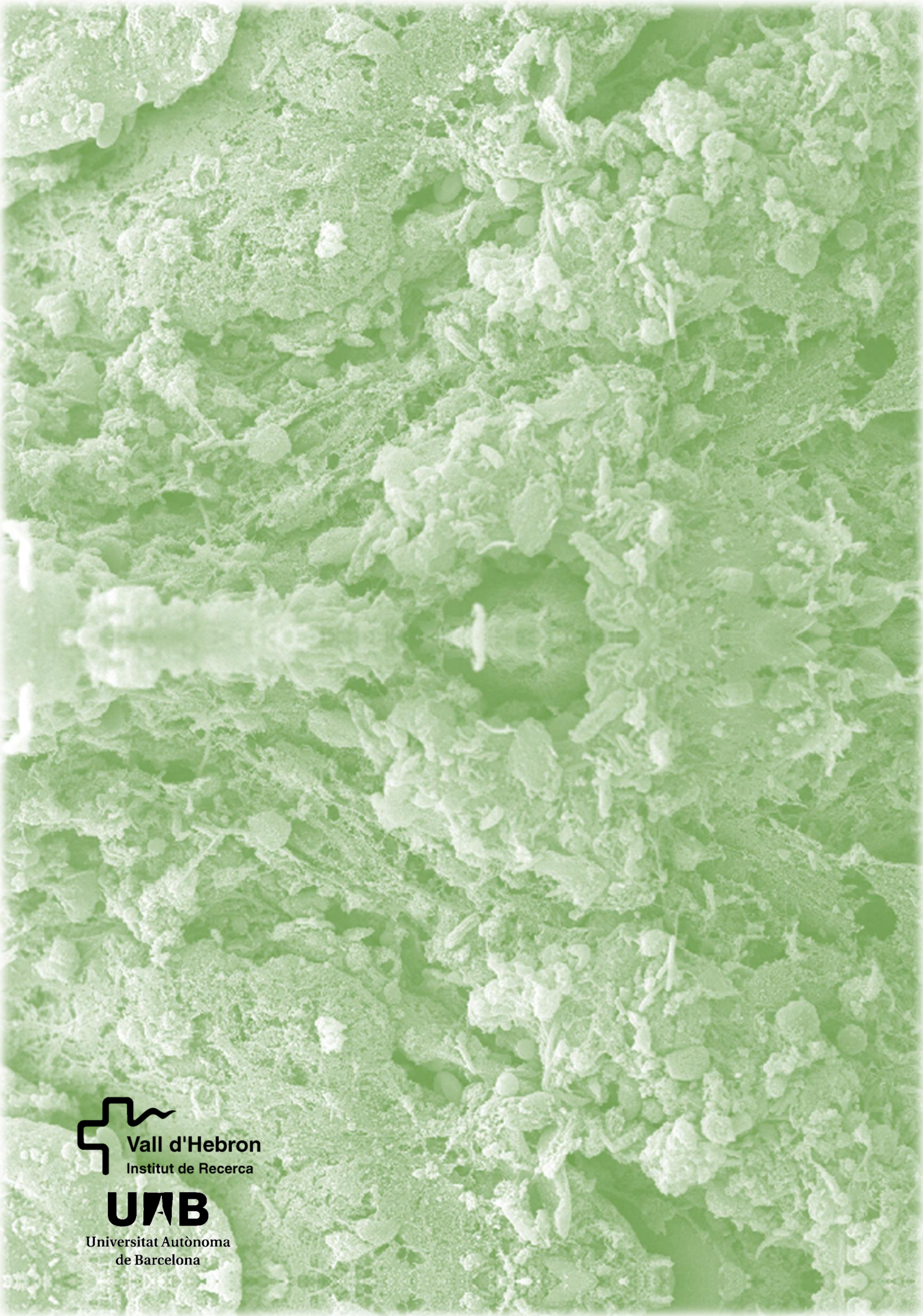
Yilmaz, B. *et al.* (2019) ‘Microbial network disturbances in relapsing refractory Crohn’s disease’, *Nature Medicine*. doi: 10.1038/s41591-018-0308-z.

Zhernakova, A. *et al.* (2008) ‘Genetic Analysis of Innate Immunity in Crohn’s Disease and Ulcerative Colitis Identifies Two Susceptibility Loci Harboring CARD9 and IL18RAP’, *American Journal of Human Genetics*. doi: 10.1016/j.ajhg.2008.03.016.

Zhong, X. *et al.* (2018) ‘Molecular and physiological roles of the adaptor protein CARD9 in immunity review-article’, *Cell Death and Disease*. doi:

10.1038/s41419-017-0084-6.

Zhong, X. *et al.* (2019) ‘The multifaceted role of CARD9 in inflammatory bowel disease’, *Journal of Cellular and Molecular Medicine*. John Wiley & Sons, Ltd (10.1111), p. jcmm.14770. doi: 10.1111/jcmm.14770.



Vall d'Hebron
Institut de Recerca

UAB
Universitat Autònoma
de Barcelona