A comparative look at different microbiome analysis pipelines

Masters thesis
by
András Asbóth
2021

Supervisor:
Eszter Ari Ph.D.
Department of Genetics, ELTE TTK

Eötvös Loránd University

Faculty of Science

Budapest

Table of Contents

I.Introduction

I/a: The human microbiome, and its significance in our well-being

The human superorganism is inhabited by around 100 trillion bacterial organisms around 90% of which are associated members of the microbiota. The relationship between human health, and microbiome constitution has long been observed, and recognized as our ‘second genome’. Our microbiome is an ecological junction heavily interwoven with various beneficial neutral and antagonistic connections, the biological epitome of ‘everything is interconnected’. (Dekaboruah et al. 2020). Bacteria in the gut carry out essential biochemical functions in our lives such as neutralizing toxins,protecting us from pathogens (mostly but not exclusively by inhibiting their colonization of our organs), breaking down complex carbohydrates (e.g. polysaccharides), and other vital, independently not digestible nutrients (Figure 1, Bäckhed et al. 2005) playing a key role in our metabolism. Some newer studies suggest our microbiome may even influence our mental health (Liu 2017).

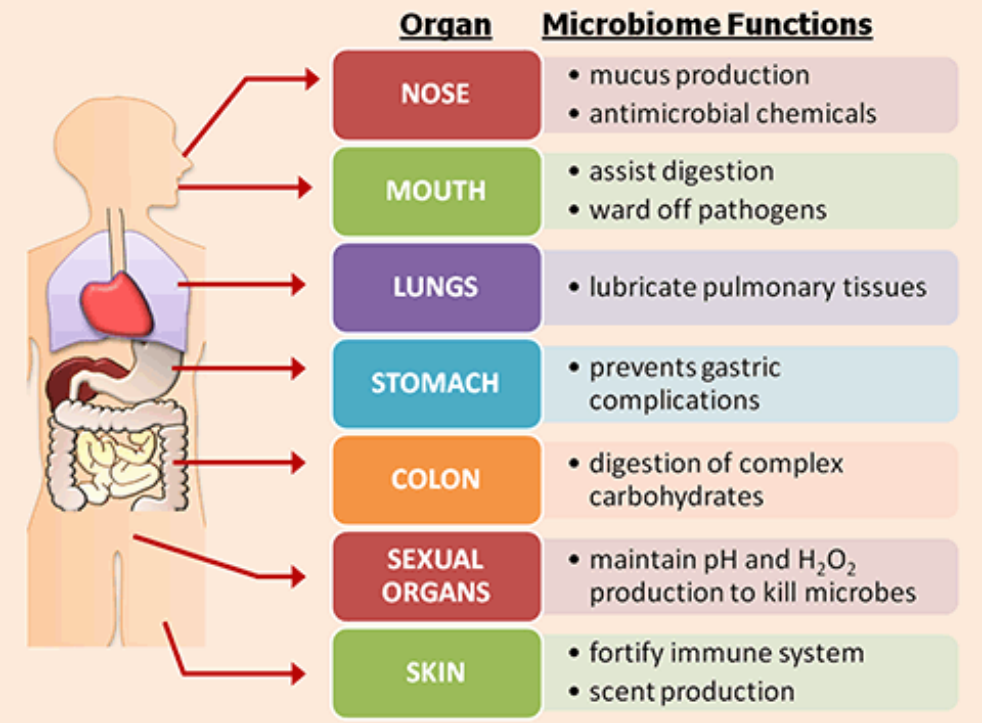


Figure 1 - Various areas in which the microbiome is present in our body, connecting them with the functions it fulfills.
(Appanna 2018)

Consequently, many of the diseases humans develop can be traced back to our body's microbiological status, clearly signifying that the bacteria in our system are associated with whether we are affected by various illnesses. A study in 2020 observed the microbiome of 3400 people, and compared its composition with the phenotypic variables of general health each person posessed (such as BMI, weight, white blood cell, and Triglyceride count, systolic blood pressure to name a few), and found significant correlations (Manor et al. 2020). As much as the gut microbiome influences our homeostasis, the outside or inner events in our body and our lifestyle also influence our gut fauna. What we eat, how much we move, seems to play a role in establishing and steering our microbiome composition (Hughes 2020, Voreades et al 2014).

I was investigating how two inflammatory bowel diseases (IBDs) and laxative treatment influence the microbial community of patients at various time points.

Inflammatory Bowel Diseases are chronic illnesses that can appear anywhere along the gastrointestinal tract. Usually they are classified into two main subtypes: Crohn's disease (CD) and ulcerative colitis (UC). Both very significant, unfortunately increasingly prevalent illnesses in the world, due to the effects of modernization, industrialization and the lifestyle changes of the last decades (Lemmens et al 2014) Both diseases are hot topics, considering the sheer number of people affected, and how much their life is inhibited (Ni et al 2017). IBD patients are frequently experiencing abdominal pain and diarrhea, with less common symptoms being anemia and weight loss (Yu and Rodriguez 2017). The non-control patients investigated were treated for either Crohn's disease and Ulcerative colitis exclusively. It has been firmly established that IBDs are associated with 'disbyosis': a swing in the composition of our gut's makeup that is different from normal. It needs to be said that this swing can only be assessed relative to the normal, as each person's baseline taxonomy levels can be very different, especially when we consider regional differences (Gupta et al 2017, He et al 2018) A direct causation of these illnesses has yet to be attributed to disbyosis itself. Abnormal microbiome may merely make a person prone to establishing these illnesses but does not directly cause them.

Crohn's disease

An IBD (inflammatory bowel disease) characterized by whole thickness transmural inflammation, which most commonly affects the ileum or the colon (Thia et al. 2008), the source location of our samples. The inflammation can be present anywhere from the mouth to the rectum. Risk factors for developing Crohn's disease include but are not limited to low fiber carbohydrate diet, and altered microbiome. With 1.5 million people affected in the US only, knowledge of a direct cause is still yet to be attained (Gajendran et al. 2018).

Ulcerative colitis

A different type of IBD, where the inflammation usually spreads from the rectum, possibly resulting in bloody diarrhea, urgency, incontinance, fatigue, irritations on the skin, and mainly more frequent bowel movements. Unlike Crohn's disease, it's usually contained in the colon, and only affects the mucosal layer of the colon, however it can also spread throughout the whole gut (Collins and Rhodes 2006, Ungaro et al. 2017).

While both IBDs sound similar in many aspects, the most dividing characteristic between Crohn's disease and Ulcerative colitis is that in CD patients experience a transmural inflammation, that can be present in any part of the gastrointestinal tract, while UC appears to cause superficial inflammation in the rectum that can possibly progressively extend into neighboring mucosa (Yu and Rodriguez 2017).

Both diseases are more prevalent in (northern) european countries and the United States (M'Koma 2013), increasingly so (Dahlhamer et al. 2016) .Both IBD-s have been associated with poor general health status (smoking, lack of exercise and sleep, major psychological distress, He et al. 2018) although there certainly is genetic predisposition (Loddo and Romano 2015), and infectious agents could also play some role (Lichtenstein 2010). Both conditions are diagnosed via colonoscopy, mainly due to lack of alternative modern methods such as biomarker profiling.

Colonoscopic investigation

Colonoscopy is a widespread inspection method of various IBDs. The inspection process is carried out via a flexible tube with a camera and sometimes a light source attached at the tip, inserted into the colon. The image of the camera is used to investigate if there are any abnormal formations (inflammations, rashes, *etc*.) in the colon area. To ensure that the bowels are empty during the process, patients are given laxatives beforehand. Consequently this causes perturbation of the gut, possibly resulting in disbyosis. The fecal samples in our dataset were obtain before administering laxatives (A), right after (B), and 4 weeks after (C). Perturbation is not the only side effect, scarring bleeding and perforation can also occur although in the minority of cases (Gevers et al. 2014). These adverse effects may also play a role in gut microbiome swings.

Significant groups of Bacteria in terms of inflammatory bowel diseases

Since 16S RNA microbiome analysis is meant to highlight how medical well-being of the gut is influenced by change of bacterial abundances, it's important to take a note on which taxa of bacteria have previously been described as influental in relation to IBD. One way to measure pro and anti inflammatory tendencies in the gut is to look at the ratio of inflammation inducing Th17 and anti inflammatory Treg cells, as they form a homeostatical system, that characterize gut tissue response (Luo et al. 2017, Schirmer et al. 2016).

Not only the Th17-Treg axis is influental however. Dysbiosis in the gut is also caused by dampening diversity in certain taxa of symbiotic bacteria, and change of abundance in others.

For instance correlation has been shown between diversity loss and severity of condition in CD patients (Gevers et al. 2014).Whether IBD occurs is down to many other factors, high gut permibility due to damaged mucosa and inflammatory lesions, low short-chain fatty acid production, low metabolic rate are all causal elements. Without going into the details of each study, (Table 1) is a conclusive collection of articles that found evidence of correlation between inflammation and abundance change.

Table 1 - Correlation between inflammation and abundance change in certain microbiota taxa. Taxa with bold letters indicate the taxa appeared in the original paper Source:(Khan et al. 2019)

Phylum	**Class**	**Order**	**Family**	**Genus**	**Effect**
Firmicutes	Bacilli	Lactobacillales	Lactobacilla-ceae	**Lactobacillus**	*Decrease*
			Strepotococca-ceae		*Increase*
		Bacillales	**Staphylococca-ceae**		*Increase*
Proteo-bacteria					*Increase*
	Gammaproteo-bacteria	Entero-bacterales	**Entero-bacteriaceae**		*Increase*
				Klebsiella	*Increase*
				Salmonella	*Increase*
				Escherichia	*Decrease*
	Gammaproteo-bacteria	Pseudomonales	Pseudomona-ceae	**Pseudomona-ceae**	*Increase*
Actinobacteria	Actinobacteria	Bifidobacterial es	Bifidobacteria-ceae	**Bifidobacteria**	*Decrease*
Bacteroidetes					*Decrease*
	Flavobacteriia	Flavo-bacteriales	Flavobacteria-ceae	**Flavo-bacterium**	*Decrease*
Firmicutes	Clostridia	Clostridiales	Ruminococca-ceae	**Faecali-bacterium**	*Decrease*
Fusobacteria	Fusobacteriia	Fusobacteriales	Fusobacteria-ceae	**Fusobacteria**	*Decrease*
Verruco-microbia					*Decrease*

I/b: Investigating the microbiome

Microbiome analysis is a highly prevalent tool in various health inspection processes, mainly due to the fact, that microbiological makeup and microbiological diversity can be a determining factor of general intestinal health, also many microbes have been associated with present conditions of various illnesses (cancer, infections, *etc.*). The fast evolution in sequencing technology provided bioinformatics with previously unfathomable amounts of data for analytical tools to process. The sequenced genetic material (RNA, cDNS) is stored digitally after it has been obtained and is used to construct sequence tables for analysis, after quality checked filtered and taxonomically identified.

The 16S region

The 16S ribosomal rRNA (Figure 2) is a 1500 basepair part of the 30S (smaller) subunit in the prokaryotic ribosome that binds to the Shine-Dalgarno sequence of the prokaryotic messenger RNA (a part of the messenger RNA ribosome binding site, next to the translation initiation codon in the 5' direction important in its identification). Since its function has remained constant through evolution, there are parts that show very reliable conservative tendencies throughout its evolution.

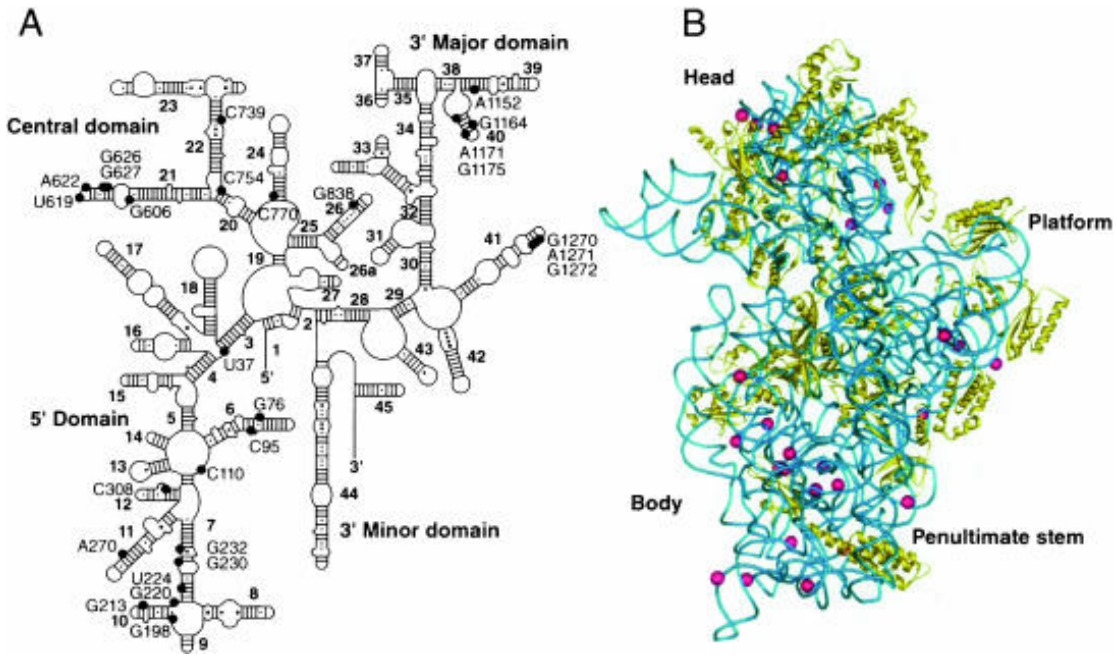


Figure 2 - The various domains in the 16S ribosomal RNA. On the left in a schematic form, marking key residuals, on the right with a structural mode. The 3' minor domain contains the anti Shine-Dalgarno sequence that facilitates the binding to the prokaryotic messenger RNA.

Source: (Ghosh and Joseph 2005)

Its consistency, and relatively short length makes it a prime candidate for usage in taxonomic classification (Clarridge 2004, Moore et al 1996, Petti et al 2005). It contains 9 notable variable regions (V1-V9) in Bacteria, each having some advantages and disadvantages when used to identify different taxa (Janda and Abbott 2007). Classification processes can take into

account one or more of these- based on the study's goals, but not all. Illumina sequencing cannot cover the whole 16S region, and long read sequencing is either less accurate or more expensive, notable examples are Minion (Tyson et al. 2018) and Pacbio (Rhoads and Au 2015).

Barcode sequences

In multiplex sequencing, unique short sequences are ligated to the ends of each library, called barcodes. To be able to pool large numbers of libraries together in sequencing for efficiency, Illumina sequencing uses barcode sequences. Pooling sequences decreases run time exponentially, and sequencers can later use the barcodes to separate libraries (in our case, samples) to get the same result as if they had sequenced them individually.

The V4 hypervariable region

Our research used the V4 hypervariable region of the 16S rRNA only. Although it is the most often used region in this type of research, it is often coupled with V3 or V6 (Bukin et al. 2019, García-López et al. 2020, Yang et al 2016).Taxonomic resolution of the V4 regions is limited to a maximum of family level (Chakravorty et al. 2007) therefore – since my analysis is also based off the V4 region, all of the results represent that.

Amplicon and Shotgun sequencing

The two widespread methods of sequencing microbial samples are amplicon and shotgun sequencing. In amplicon sequencing a specific region is selected via specific primers to sequence a chosen gene or gene fragment, such as the 16S region in the case of Bacteria, and Archaea. With the shotgun method, the entire DNA content of the sample is amplified via random primers. Amplicon sequencing can offer a general look at the microbial makeup, that is very cost effective, but the sequencing will suffer from PCR bias the effect of intrinsic differences between primers, and late self annealing in most abundant primers (Acinas et al. 2005). Also no data is gained besides the selected gene (such as the properties of the host species). Another problem for instance, in the 16S case, is that the specific gene can appear more than once in the genome (making it seem like there are multiple occurences), and can even have different copies in the same genome. Shotgun metagenomics provide a more detailed whole genom analysis that can circumvent the problems above, and since not only a single selected gene fragment is amplified, functional information about the rest of the genes

can also be obtained. It's drawbacks are that it is more complex to analyse and much more expensive. Our study used the option of amplicon sequencing, because it is more cost effective, and in the first stage of research a general look was enough. Additionally, the questions we investigated did not require functional information to be answered.

Reference databases

When analysing a new sample, the algorithms compare the newly sequenced DNA or RNA to previously collected and verified ones. Every day scientists are annotating sequences, which are then collected into databases that are used for categorization work. In the case of the 16S ribosomal RNA genes, the largest and most well annotated, updated databases are RDP and the SILVA. For this reason, the choice of database is crucial. Low accuracy train sets can not only fail to identify lower levels of identity (genus, species for instance) but can actually misidentify the sequence. Obviously, the more variable regions we consider, the better our accuracy will be, regardless of train sets (in most cases). Various articles suggest, that different combinations of variable regions are more suited to identifying different bacteria (Bukin et al. 2019, Chakravorty et al. 2007, Fuks et al. 2018), but researchers still strive to find an optimum between using too much of them, but enough to make the results significant as using more variable regions is more resource consuming.

I/c: Survey of IBD patients

Properties of the dataset

My dataset consisted of the fecal samples from measurements undertaken by **Mariann Rutka Ph.D. First Department of Medicine, University of Szeged, Faculty of Medicine**, operating in the group of **Professor Tamás Molnár**. The ill subjects were treated in hospitals for the two aforementioned IBDs, the rest of the subjects were in the control group. Patients from all three groups had their samples taken, *before*, *right after*, and *4 weeks* taking laxatives, for the purpose of colonoscopic examination, corresponding to the A, B and C time points in the metadata. The patients were grouped based on their condition, into UC (Ulcerative colitis), CD (Crohn's disease), and control groups, resulting in a total of 41 patients, consisting of 9 CD 13 UC and 19 control participants, concluding in 123 samples with one forward and one

reverse read of the 16S V4 loop, in the form of *fastq* files. The sequencing was outsourced to Seqomics Kft, Their report on the procedure is quoted here:

"Determination of bacterial microbiome component: Faecal microbiota structure was determined by sequencing the V4 hypervariable region of the 16S rRNA gene. The metagenomic DNA was extracted from frozen stool samples by using ZR Faecal DNA MiniPrep™ kit (Zymo Research) following the instructions of the manufacturer. For DNA isolation we needed ≤ 150 mg faecal sample/individual. The V4 region of the 16S rRNA gene was amplified with indexed Illumina primer-pairs (i5-i7) using Phusion high-fidelity DNA-polymerase (Thermo Scientific). In the PCR mix the final concentration of the primers was 0.4 μM and the template DNA was used in 20 ng concentration. PCR products were isolated from gel and after purification sent for dual-index paired-end Illumina MiSeq sequencing (250 bp reads)."

II. Methods

II/a: The applied pipelines:

Mothur and *dada2* are the two pipelines I'm going to be contrasting in my thesis. The product of both are the count tables that are the input of the post classification analysis (Alpha diversity, Beta diversity and abundance comparisons). I applied the two pipelines from the 'importing and merging reads – initial filtering' to the 'Removing non relevant taxa' steps. Demultiplexing was done by Seqomics Kft.

Dada2 is an *R* (or *Qiime*) integrated package, that relies on a few other *R* packages (most importantly *Biostrings, ggplot2, reshape2, Rcpp*) to work. It is handled in *RStudio* which makes it very accessible to scientists, who already probably have the IDE installed. *Mothur* is a C++ based command line tool, that requires a bit more flexibility on the user's part, and relies less on visualizations as it does not use a graphical user interface. It should also be noted that since *dada2* uses ASVs instead of OTUs it is considered a state of the art classifier, while *Mothur* is an older, tried and true software, currently boasting 15339 citations (Schloss et al. 2009).

II/b: Processing sequence data:

Demultiplexing reads

To decide which read belongs to which sample, they have to go through a process called demultiplexing, during which each read is sorted into a sample based off its sample associated barcode. In Illumina sequencing, a few bases are attached to the reads of each sample for identification. These bases are read by the computer, it identifies the sample and inserts the read into the appropriate *fasta* file.

Importing and merging reads – initial filtering

After importing the reads, the user receives a plot quality profile, either visually (Figure 3) or as a form of chart, that provides an aid when selecting a cutoff point for filtering. In Illumina sequencing, quality scores are usually diminishing towards the end, which is where the cutoff point is mostly made. Quality is measured by scoring bases using the Phred quality score system (Ewing et al. 1998).

Phred scores

Phred score is described with the following equation:

$Q = -10/log\{10\}P$

where P is the Phred score and Q is the probability of a base being incorrect. For instance if the score is 20 (towards the end of the reads), P is 0.01 which means the probability of the base being wrong is 1% (Ewing et al. 1998).

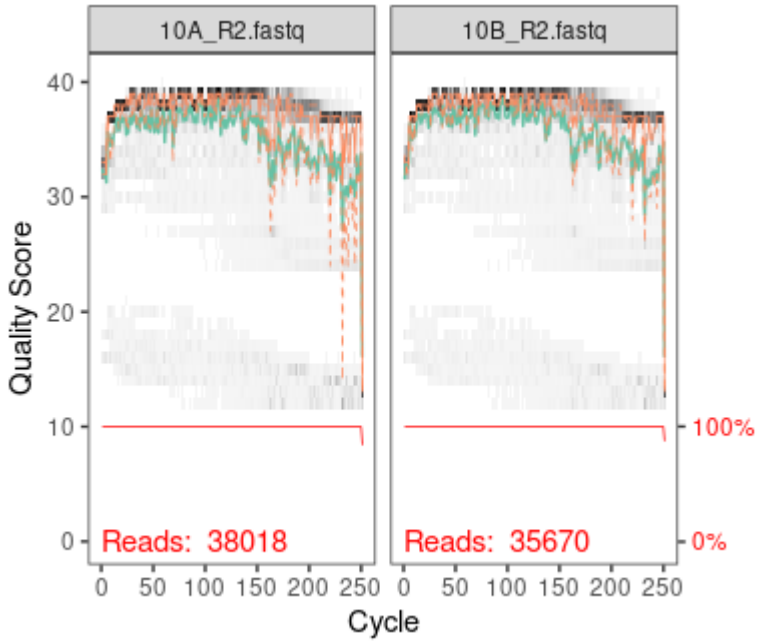


Figure 3 - Dada2's Phred quality profile chart, representing quality scores along the read. At around the 200th base the score starts dropping.

Source: Own analysis

Sequence data is imported as *fastq* files with each sample having two reads, one forward one backward, usually signified as *R1* and *R2* respectively. During the merging step of the analysis each forward and reverse read is merged into a single fragment, based on their IDs and overlapping sequences. Included in this step, there is an initial filtering, and discrepancies between the two reads are removed. At the end of this step the *fastq* files are converted into the analysing software's dataformat to be handled in the pipeline.

Mock samples

Most analysis pipelines also advocate for the inclusion of a mock sample to check for the legitimity of reads (Pollock et al. 2018),I decided included the one offered on their website. mock samples contain predetermined amounts of bacterial sequences, whose microbial makeup is manually prepared, they are used in the pipeline to verify if the classifier works correctly.

Removing chimeras

After the sequences were clustered, filtering is needed for chimeric sequences. Chimeras are hybrids created by a combination of multiple sequences. The error that causes their appearance occurs during the PCR amplification and the resulting reads are compromising the further analysis. They are identified by checking whether they can be reconstructed by combining 2 or more other sequences. My samples were quite high in chimeras with them making up more than half of the total number, therefore this is a very crucial step in the pipeline.

Constructing the abundance table

After the sequences have been filtered and assigned to samples, the next step is the creation of a table that represents how much of each different sequence is available in each sample. To solve this problem, two approaches exist: We can cluster sequences either into Operational Taxonomic Units (OTUs) or (Amplicon Sequence Variants) ASVs. The two concepts are similar, but not exactly the same: OTUs are a batch of sequences grouped based on a cutoff similarity level. The level of specificity can be very different (from species to genus and family levels) and usually depends on the accuracy of sequencing, variable regions used and the level of similarity we apply to create the OTUs (usually 99%). Generally the smaller taxonomic units researchers try to classify, the bigger the computational demand. ASVs on the other hand are differentiated by only as much as a single nucleotide change, making them much more precise in resolution. This is not the only advantage over OTUs, ASV can be much more universal, if they are used in many different type of studies. The exact sequence amplicon can translate much easier into another pipeline than a taxonomic classification (Callahan et al 2017, Porter and Hajibabaei 2018). *Dada2* is an ASV based classfier, praised for it's accuracy (Prodan et al. 2020) meaning that it utilizes the resolution of each unique sequence, while *Mothur* is OTU based, and only differentiates until reaching a certain similarity.

While every sequence could individually be represented in the abundance tables, it would not be the most consequent thing to do by default, because we cannot tell whether the sequences are unidentical because of a real taxonomic dissimilarity, or because of a sequencing error. The solution of *Mothur* (OTU-based) is to gather all sequences based on a percent similarity into one OTU based off the philosophy of considering a low amount of errors to be random,

but once there are more than the percent threshold, it means the sequence is part of a different OTU. In *dada2* a machine learning is used to learn to differentiate between sequencing errors and base changes. Theoretically this would mean that only truly difference sequences are kept as separate ASVs (of course no machine learning algorithm is perfect).

As a next step of the analysis we create and abundance table (Table 2) by merging individual sequences into OTUs or ASVs. In an abundance table, each OTU's/ASV's abundance is shown in each sample.

Table 2 - The layout of a typical count table. Columns are representing the samples, the row names contain the OTUs or ASVs.

	Sample 1	Sample 2	Sample 3	Sample 4...
OTU1/ASV1				
OTU2/ASV2				
OTU3/ASV3				

Taxonomic classification by alignment

Assigning taxonomy most of the times is computationally the longest step in the pipeline, that also requires the most processing power, as even with a highly optimized workflow sequence alignment requires high amounts of underlying computation (Wang et al. 2007). To begin sorting the sequences, a train set is used on which the classifier algorithm can calibrate itself to match the ASVs/OTUs into taxonomic categories. I chose the most up do date version of version of the SILVA database's species train set 138.1 updated on March 10 2021. The mock /reference files are also crucial in this step because the learning algorithm's results are validated using its artificial abundance makeup.

Removing non-relevant taxa

There are always various taxa in the samples, that we want to exclude from our analysis. Since our sequencing used bacterial 16S rRNA specific primers, it is advised to remove Archaea and Eukaryota, since they are surely sequencing errors, leaving only Bacteria included. The naming of these groups is dependent on the dataset we use for classification, so I proceeded to change them and use the appropriate ones for both EzBioCloud and SILVA.

II/c: Post classification statistical analysis:

Introduction

After classifying our samples, the mains steps most investigations take are measuring Alpha diversities, that is the absolute diversity of the samples, Beta diversities, a metric for the relative diversities between the samples, and abundance changes of taxonomic units. All of these comparisons are meaningful in the light of the knowledge we have about the real life conditions of our samples -which groups of patients they belong to, or at which time point they were taken.

Alpha diversity

Alpha diversity measures the diversity within a single sample. This results in an absolute, stand-alone value, that signifies the given sample's degree of diversity. It's important to note however, that this score doesn't tell us anything about the makeup of the sample in the sense that it doesn't evaluate scores based on what kind of taxa it finds. Theoretically it could have happened that two entirely differently composed samples can have the same Alpha diversity, and not have any bacteria in common. To measure Alpha diversity, I used the most commonly applied Shannon diversity index:

$$H=-\Sigma p_i \log(b) p_i$$

In the above formula H is the resulting value p_i is the proportional abundance of the species i and b is the base of the logarithm. The theoretical maximum of the Shannon index is achieved if all of a sample contains the same level of all taxa (only happening in theory) while the maximum is in the scenario when all of the bacteria are from a singular taxa.

To visualize the Alpha diversities I plotted each individual patient's Alpha diversity corresponding to his or her time points in a line plot, to see how it has changed between measurement points. I also plotted the Alpha diversities of patient groups using boxplots, for instance to see if healthy people in general have a more diverse microbiome or not. The diversity metrics was implemented in *R* using the *vegan* (Oksanen et al. 2020) package's **diversity**, function, in which you can select from various diversity calculation indexes.

Beta diversity

Beta diversity measures how *different* the diversity compositions are between each sample. Using this metric we can look up how much the microbiome *changed* after the administration of laxatives between IBD and control patients. or how different in general are the different patient groups or times. We could also compare if healthy or sick groups responded with more or less change in their bacterial makeup, following the medical intervention. Two notable methods of calculating Beta diversity are the Unifrac and the Bray-Curtis models. I implemented the Bray-Curtis diversity metric in *R* using the *vegan* package's vegdist, function, in which you can select from various diversity calculation indexes (Oksanen et al. 2020). I ran the Bray-Curtis distance measuring method. The resulting matrix can be visualized using principal coordinate analysis, detailed next.

Bray-Curtis distance

$$d_{jk}=(\sum|(x_{ij}-x_{ik})|)\ /\ (\sum(x_{ij}+x_{ik}))$$

In the formula, x_{ij} and x_{ik} refer to the abundance of species (column) i and sites (rows) j and k. Note that Bray-Curtis distance (or any other Beta distance for that matter) (d_{jk}) can only be interpreted relative to another sample (unlike Alpha diversities). In the produced Bray-Curtis matrix, each sample has a relative distance to each other sample. This chaos can be efficiently resolved using different principal coordinate analysis methods later on (see discussion). After that step, I plotted the samples in various groupings, using a traditional scatter plot, to see if there are any outliers, and to take a general look at how the distribution looks.

Principal Coordinate Analysis

The first step of Principal Coordinate Analysis (PCoA) is to measure the distances between the samples (Bray-Curtis distances, calculated with *vegan*). In the case of microbiome analysis, the distances between different taxa create so many axes, that it cannot possibly be explained by a two dimensional diagram. The way PCoA aims to solve this issue is by creating new coordinates by performing an eigenanalysis (also called single value decomposition). These are then ranked based on their eigenvalues, the score rating how many percents of the variation a principal coordinate (PC) is able to explain. If the algorithm manages to find for instance a PC that explains 36 percent and another which explains 15

percent of the variance, the investigation can present a 2D graph that, in total explains 51% of the entire distance between samples (Jolliffe 2002).

Differential abundance analysis

In differential abundance analysis we try to pinpoint those taxa which had significantly increased or decreased counts between two samples or groups of samples. The goal of this measurement is to reveal if any taxa has significantly changed in between sample groups. For instance if in Con-B there is less of a certain bacteria than in Con-A we could infer that the effect of the laxative treatment was detrimental on the taxon. *DeSeq* (Love et al 2014) and *edgeR* (Robinson et al 2010) are two of the more popular differential abundance analysis tools, for analysing RNA-Seq, ChIP-Seq or HiC abundance data. Both of them are integrated into *R* and provide throughout documentation of their functions (Varet et al. 2016). Their normalization methods have been shown to perform above their peers on RNA data (Dillies et al. 2013). I picked *edgeR* mainly because I was already familiar with it, and used the TMM (Trimmed-mean M values) normalization method (Robinson and Oshlack 2010) since it is a robust application able to account well for the library size and read depth bias between the samples. Abundance change in *edgeR* is measured by the log fold change parameter (*logFC*), which represents ratios of the normalized count values of the compared bacteria. It means that the change of abundance was two to the power of the *logFC* number between the samples. To give statistical meaning to abundance changes an adjusted *p-value* is calculated for each change (via the Benjamini-Hochberg procedure in my case), to make apparent which change is probably significant and not due to random chance.

III. Results

Introduction

I ran the *R*-based *dada2* package and the *C++* implemented *Mothur* pipeline on our dataset, using the same 16S rRNA database, which for this purpose was the up to date version of SILVA. I used the same mock sample in evaluation of both taxonomy assignments. After the count table was ready, I imported it into *R* and aside from basic formatting, that was needed to make the count tables uniform in terms of indexes and columns, I followed the same analysis steps, measuring Alpha and Beta diversities, and measuring abundance differences between groups.

Abundance tables

Here I present a few statistics regarding both tables (Table 3), to provide a quick look at the differences between the pipeline results.

Table 3 - Statistics of the abundance tables resuling from the two analysis pipelines

Source: Own analysis

	Dada2	***Mothur***
mean sample count sum	40332.41	41619.46
maximum sample count sum	108480	112448
minimum sample count sum	8426	8583
standard deviance in sample count	18109.17	18670.72
Dada2 ASVs / Mothur OTUs	70	137

Both samples are very close in count numbers, we can only see minor differences between their values, except for the number of ASVs and OTUs.

Next I demonstrate the usage of different sequence analysis tools to construct abundance tables, looking at their individual advantages and drawbacks when used for further analysis,

and the differences between the resulting abundances of bacteria. I'm also aiming to highlight some of the advantageous various features that are unique to the pipelines I used.

III/a: Results of the pipelines:

Alpha diversities

I visualized each person's line plots separately. (Figure 4) In these figures, an individual patient's gut- diversity is shown over time (A, B, C) time points. It is especially useful to see line plots separated by patients, when we want to examine treated personnel individually. The two analysis tools produce so similar results, that difference can barely be seen especially on the first ~25 samples. In the next pictures, I drew *dada2* with blue and *Mothur* with red, to highlight differences.

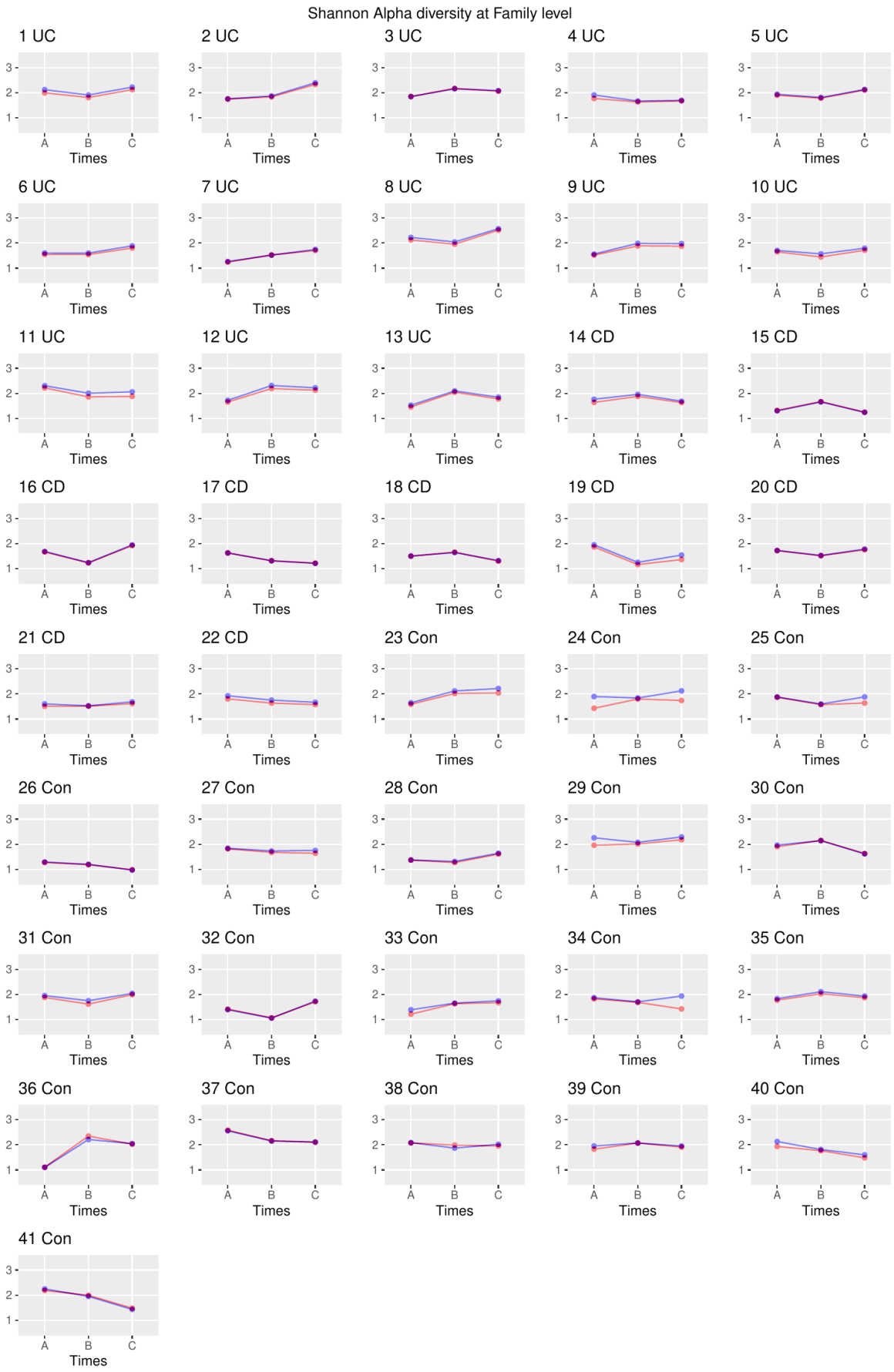
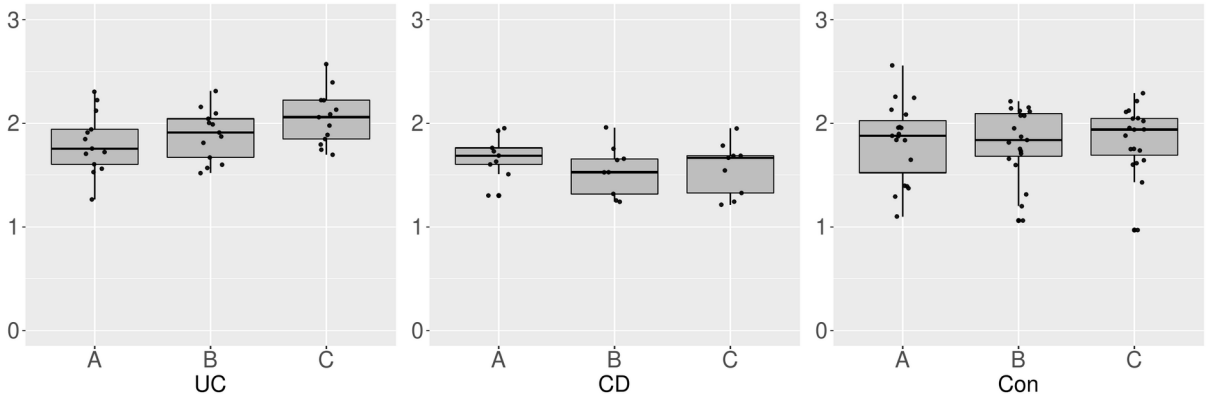

Shannon Alpha diversity at Family level
1 UC
2 UC
3 UC
4 UC
5 UC
6 UC
7 UC
8 UC
9 UC
10 UC
11 UC
12 UC
13 UC
14 CD
15 CD
16 CD
17 CD
18 CD
19 CD
20 CD
21 CD
22 CD
23 Con
24 Con
25 Con
26 Con
27 Con
28 Con
29 Con
30 Con
31 Con
32 Con
33 Con
34 Con
35 Con
36 Con
37 Con
38 Con
39 Con
40 Con
41 Con
Times
A
B
C
1
2
3

Figure 4 – The alpha diversities of the 41 patients.The values of the dada2 and Mothur pipelines were drawn with blue and red respectively. On the y axis Alpha diversity values are shown, the x axis represents the time points when samples were taken. In most the cases everything appears purple, because dada2 and Mothur produce essentially the same results. Although, in some samples, for instance in sample 28,29,3031, and 40 we can see some differences.

Source: My Analysis

I also measured the Shannon alpha diversities for all samples separately on the family level and visualized the results on boxplots (Figure 5).

Dada2:



Mothur:

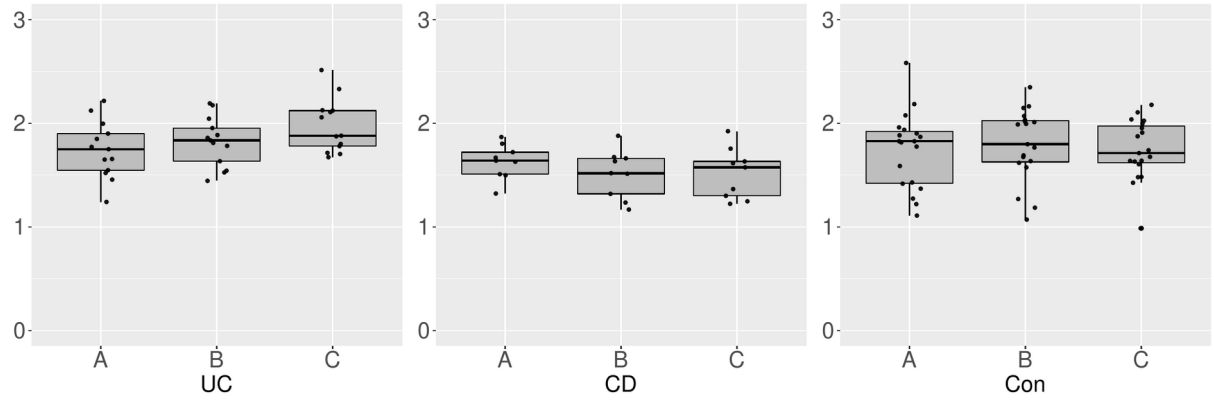


Figure 5 - Alpha diversities for the various diagnosis groups and their time points, calculated by dada2 (up) and Mothur (down), visualized as boxplots on a scale of 0 to 3. The jittered points represent individual sample. The middle, bold line in a boxplot represents the median of the group, the ends of the box the first and third quartile, while the end of the vertical lines represent the lowest and highest points in the group.

Source: My analysis

On the boxplots, (Figure 5) we can see a steady increase in the samples of UC patients, suggesting that their overall microbiome diversity was increasing during treatment, while CD is generally lower, has less of a deviance, while control seems to be all over the place.

Table 4-5 represents all authoritative comparisons of Alpha diversities, using both pipelines:

Table 4 – The alpha diversity comparisons in dada 2. Significantly different Alpha diversity comparisons are highlighted with green.

Source: Own analysis

dada2	Group	α	Group	α	P-val
	CD-A	1.67	Con-A	1.82	0.17
	UC-A	1.8	CD-A	1.67	0.39
	UC-A	1.8	Con-A	1.82	0.7
	CD-B	**1.54**	**Con-B**	**1.81**	**0.02**
	UC-B	**1.88**	**CD-B**	**1.54**	**0**
	UC-B	1.88	Con-B	1.81	0.76
	CD-C	**1.56**	**Con-C**	**1.84**	**0.02**
	CD-C	**1.56**	**UC-C**	**2.04**	**0**
	UC-C	2.04	Con-C	1.84	0.08
	CD-A	1.67	CD-B	1.54	0.29
	CD-A	1.67	CD-C	1.56	0.38
	CD-B	1.54	CD-C	1.56	0.66
	Con-A	1.82	Con-B	1.81	0.77
	Con-A	1.82	Con-C	1.84	0.86
	Con-B	1.81	Con-C	1.84	0.88
	UC-A	1.8	UC-B	1.88	0.51
	UC-A	1.8	UC-C	2.04	0.05
	UC-B	1.88	UC-C	2.04	0.18

Table 5 – The alpha diversity comparisons in Mothur. Significantly different Alpha diversity comparisons are highlighted with green.

Source: Own analysis

Mothur	Group	α	Group	α	P-val
	CD-A	1.62	Con-A	1.73	0.26
	UC-A	1.74	CD-A	1.62	0.29
	UC-A	1.74	Con-A	1.73	1
	CD-B	**1.51**	**Con-B**	**1.79**	**0.02**
	UC-B	**1.82**	**CD-B**	**1.51**	**0.01**
	UC-B	1.82	Con-B	1.79	1
	CD-C	**1.51**	**Con-C**	**1.74**	**0.03**
	CD-C	**1.51**	**UC-C**	**1.97**	**0**
	UC-C	**1.97**	**Con-C**	**1.74**	**0.02**
	CD-A	1.62	CD-B	1.51	0.38
	CD-A	1.62	CD-C	1.51	0.25
	CD-B	1.51	CD-C	1.51	1
	Con-A	1.73	Con-B	1.79	0.58
	Con-A	1.73	Con-C	1.74	0.81
	Con-B	1.79	Con-C	1.74	0.52
	UC-A	1.74	UC-B	1.82	0.51
	UC-A	**1.74**	**UC-C**	**1.97**	**0.04**
	UC-B	1.82	UC-C	1.97	0.26

This seems to reinforce that UC's Beta diversity is increasing, as seen on the boxplots. The UC-C sample it's significantly higher than the original diversity (UC-A), and both Con-C and CD-C. Overall the CD group's alpha diversity is also significantly smaller than the other two. I created a collage of sample group pairs that were significantly different from each other in the analysis (Figure 6-7):

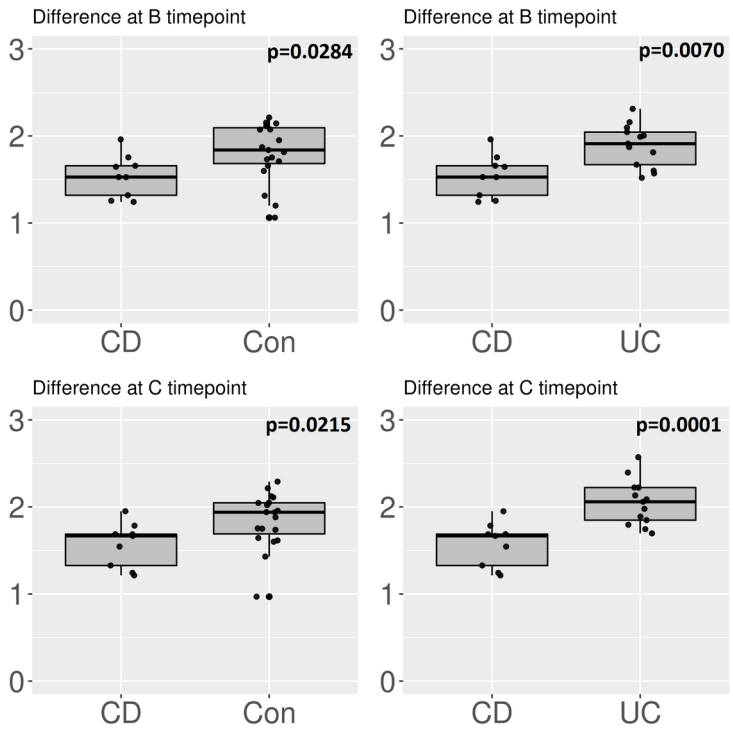


Figure 6 - Significant differences in Alpha diversities at time points corresponding to the title of the subplots – dada2.

Source: Own analysis

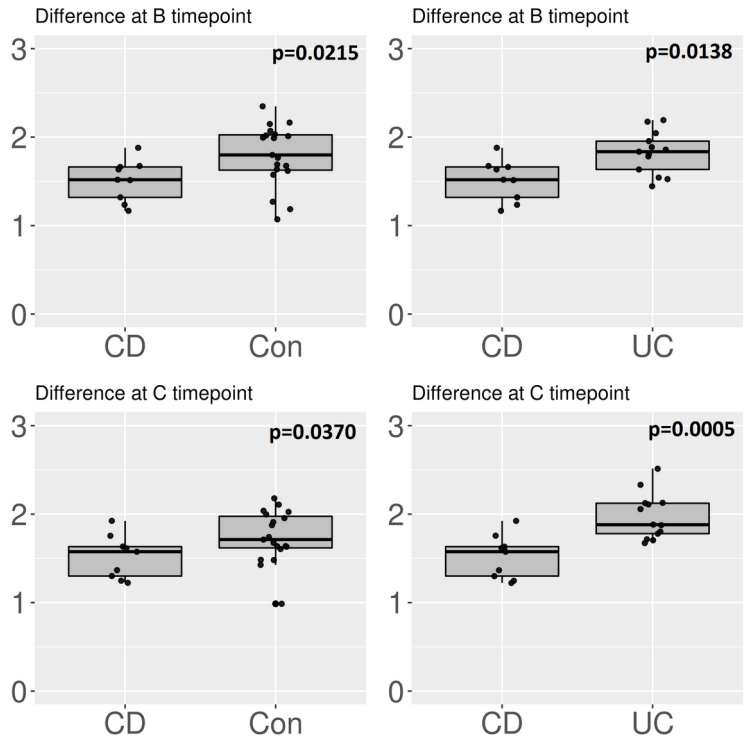


Figure 7 - Significant differences in Alpha diversities at time points corresponding to the title of the subplots – Mothur.

Source: Own analysis

Findings exclusive to *Mothur*

There were two differences in Alpha diversity, that only the *Mothur* pipeline deemed significant. Based on that abundance table, UC was significantly higher than Con group at time point C, and UC patients increased from time point A to C. Figure 8 represents these changes:

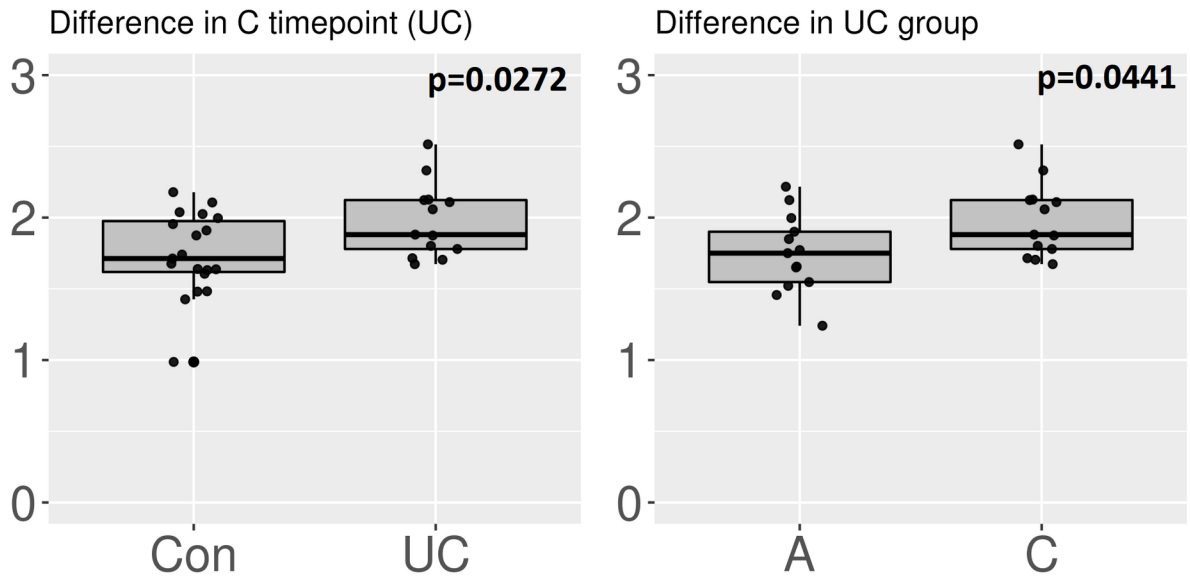


Figure 8 – Significant differences between UC and Con at time point C (left), and between UC at time point A and C (right).

Source: Own analysis

Beta diversities

The vegdist function takes the count table as an input, to see how many of each bacteria families were present in each sample, the calculated Beta diversity matrix is the input of the PCoA process I described in the methods section. The Principal Coordinates Analysis tries to find the best coordinates that can explain the distances between the samples the most (Figure 9 and 10).

One of the leading questions of my thesis was whether the microbiome of IBD patients was more disturbed laxatives, and how well it could recover from it. To measure this, I ran Mann-

Whitney-U tests between the Beta diversity values of groups, comparing the following metrics:

- A-B Beta distances
- A-C Beta distances
- A-B Beta distances divided by A-C Beta distances (the higher the better)

The resulting values (mean of all Beta distances in the subgroup) of both pipelines were very similar, which I present here in a table (Table 6).

Table 6 - The mean values of distances, in the groups (rows) between the time points (columns)

Dada2	**A -C**	**A -B**	**A -B / A -C**		**Mothur**	**A -C**	**A -B**	**A -B / A -C**
UC	0.37	0.37	1.19		**UC**	0.36	0.35	1.23
CD	0.41	0.39	1.04		**CD**	0.39	0.38	1.02
Con	0.43	0.42	1.1		**Con**	0.42	0.42	1.12

In the next table (Table 7) I list all the meaningful comparisons I've made with the results using *dada2* and *Mothur*.

Table 7 – Listing the comparisons, mean Beta distances of the first and second group of comparison, and the p-value of the Mann Whitney test. The higher values in significant pairs (green) are bold.

Source: Own analysis

Dada2	**1st**	**2nd**	**P-value**	**Mothur**	**1st**	**2nd**	**P-value**
UCA - UCB CDA - CDB	0.5	0.49	0.38	UCA - UCB CDA - CDB	0.5	0.49	0.58
UCA - UCB ConA - ConB	0.5	0.53	0.14	UCA - UCB **ConA - ConB**	**0.5**	**0.53**	**0.03**
CDA - CDB **ConA - ConB**	**0.49**	**0.53**	**0.03**	CDA - CDB **ConA - ConB**	**0.49**	**0.53**	**0.03**
UCA – UCC CDA - CDC	**0.53**	**0.48**	**0.03**	**UCA - UCC** CDA - CDC	**0.53**	**0.48**	**0.01**
UCA - UCC ConA - ConC	0.5	0.52	0.27	UCA - UCC ConA - ConC	0.49	0.52	0.2
CDA - CDC ConA - ConC	0.48	0.52	0.1	CDA - CDC ConA - ConC	0.48	0.52	0.11
UCA - ConA UCB - ConB	0.54	0.53	0.35	UCA - ConA UCB - ConB	0.53	0.52	0.4
UCA - ConA UCC - ConC	0.54	0.55	0.46	UCA - ConA UCC - ConC	0.53	0.55	0.26
UCB - ConB UCC - ConC	0.53	0.55	0.1	UCB - ConB UCC - ConC	0.52	0.55	0.06
CDA - ConA CDB - ConB	0.55	0.55	0.46	CDA - ConA CDB - ConB	0.55	0.55	0.45
CDA - ConA CDC - ConC	0.55	0.55	0.82	CDA - ConA CDC - ConC	0.55	0.56	0.66
CDB - ConB CDC - ConC	0.55	0.55	0.33	CDB - ConB CDC - ConC	0.55	0.56	0.19

The Beta distance matrix is the input of the PCoA process I described in the methods section. The Principal Coordinates Analysis tries to find the best new artificial variables that can explain the distances between the samples the most (Figure 9)

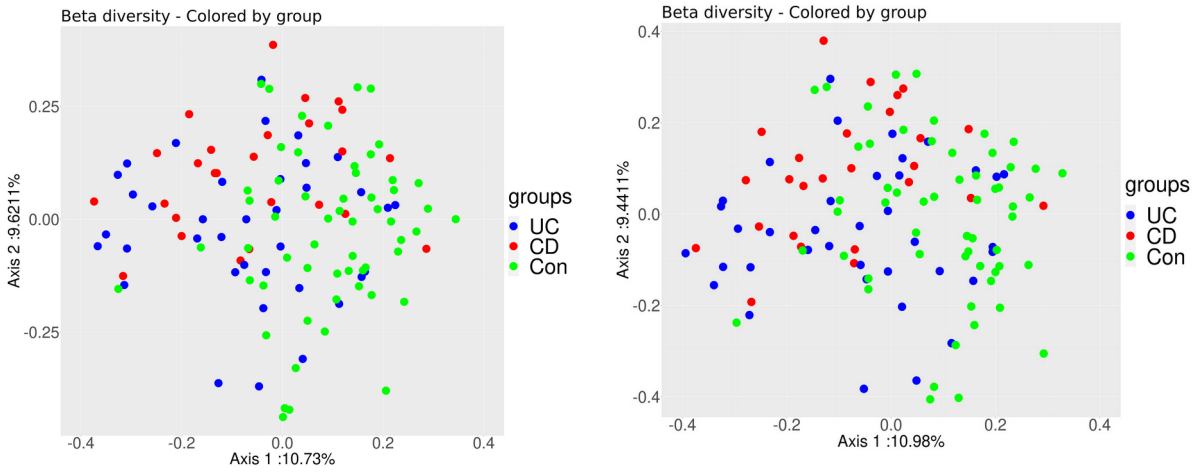


Figure 9 - Plots of the Beta diversity PCoA. dada2 left, Mothur right. The groups are based on treatment groups, UC CD and Con. On the axes, the eigenvalues are shown as the percentage of the variation they explain.

Source: Own analysis

Again the two pipelines produced very similar plots, but without reasonable separation between the groups. In a general sense I conclude that PCoA did not find significant difference because of the low value of the eigenvectors. The entire plot together explains around 20% of the variance which is considered low.

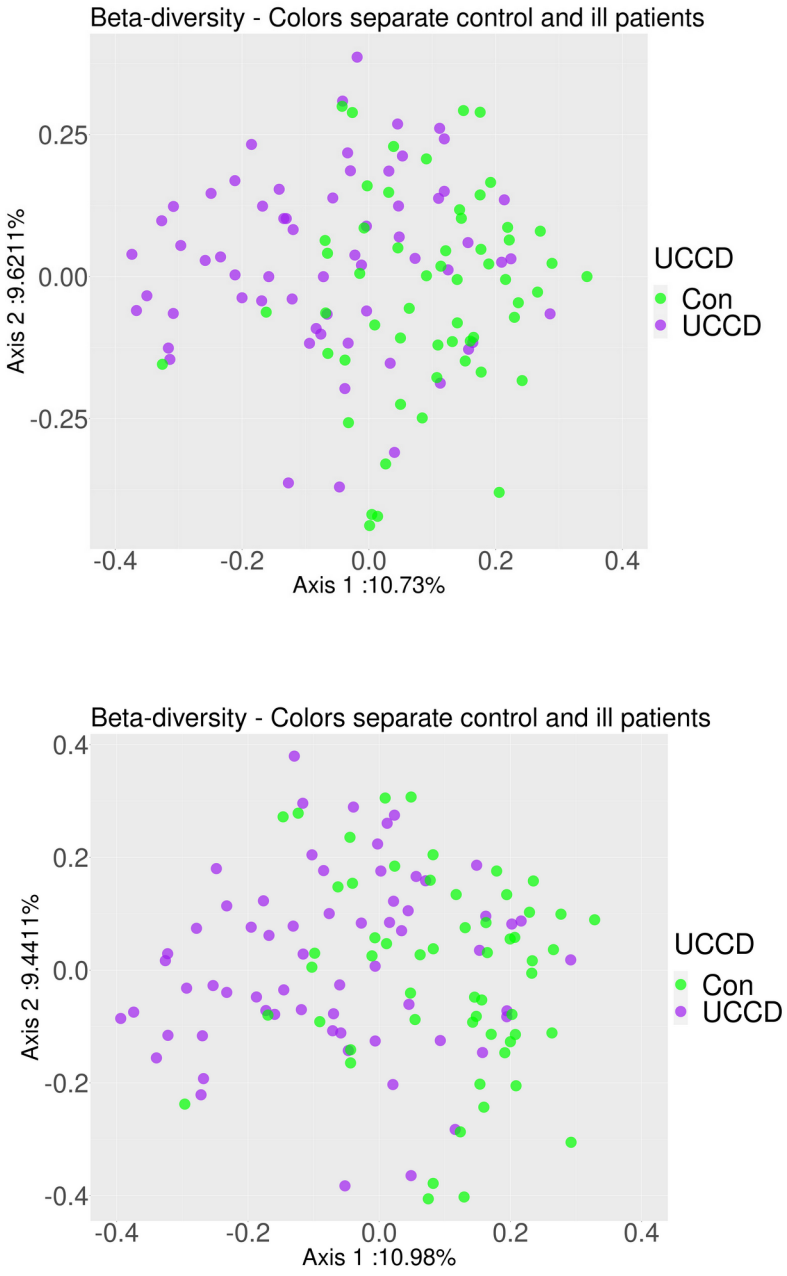


Figure 10 - Plots of the Beta diversity PCoA. dada2 above, Mothur below. The groups are based on treatment groups, UC+CD and Con. On the axes, the eigenvalues are shown as the percentage of the variation they explain.

Source: Own analysis

Although it's not significant, I still observed some separation in Beta diversity between the control and the ill group of patients on the PCoA plot (Figure 10).

Differential abundance analysis

The last part of the analysis is investigating the differential abundance between groups and time points. This was done via the *edgeR R* package (Robinson, McCarthy, and Smyth 2010). I compared every reasonable combination of groups and times, and selected the bacterial families that have significantly changed during observations. I elaborated on those ones that are said to be associated with IBDs.

Streptococcaceae

Dada2 has found that the abundance of Streptococcaceae was significantly higher in the B time point of the CD group than the B time point in the control sample. (*dada2*: 8689 -> 8885).

Clostridiaceae and Enterobacteriaceae

The Mothur pipeline has suggested that Clostridia is higher in control, when we consider all the samples together (regardless of time points).

From A to B in the control group

From time point A to B (before and after taking laxatives) the control group had the only taxonomy changes (*dada2*: 4, *Mothur:* 8). The hypothesis I can attribute to this phenomenon is that those suffering from IBDs already experience diarrhea on a fairly regular basis, so the taxa that are sensitive to laxatives have already been eliminated and more enduring ones have taken their place, while in the control group they are still present. Another observation is that some of the most growing families Brucellaceae, Moraxellaceae, and Alcaligenaceae all contain various opportunistic pathogens.

Individual outbreak - Pseudomonaceae

The fold change analysis based on both count tables suggest, that both the UC and CD group has less of the taxa Pseudomonaceae than Con, in both pipelines in points B and C. This seems to contradict literature, which suggests, that Pseudomonaceae may play a role in IBD and that it increases when a patient establishes IBD (Scaldaferri et al. 2013, Wei et al. 2002).

Upon closer inspection however, I realized that the difference is due to two individual outbreaks in patient 37 B time point and 41 C time point who probably had an independent factor involved in the fold change. They had counts of 826 (*Mothur*)/829 (*dada2*) and 32746 (*Mothur*)/32676 (*dada2*) respectively. It was interesting to see that in time point B Patient 41 already had the second highest amounts of Pseudomonaceae after Patient 37 with (*Mothur*) 11/(*dada2*) 12, so these bacteria might have been the precursors to the multiplication later.

In conclusion, while some observations correlated with literature, about equal amounts didn't. I believe the 41 patients in this study, among whom only 9 was representing the CD group for instance (13 UC and 19 Con) it is very problematic to draw so specific conclusions as what type of Bacteria increased from time point to time point. One thing the abundance analysis further confirms is that the identified amounts of taxa were also very similar in the results of *Mothur* and *dada2*.

Total abundances

In my analysis I also compared the ten most abundant families in each time and group. While this comparison can provide a general look at the abundance, it should be used with caution, understanding that since some bacteria can contain multiple 16S domains, and individual outbreaks of bacteria can mean very large read numbers therefore the table's data can be skewed that effect (Figure 11-12).

Dada2

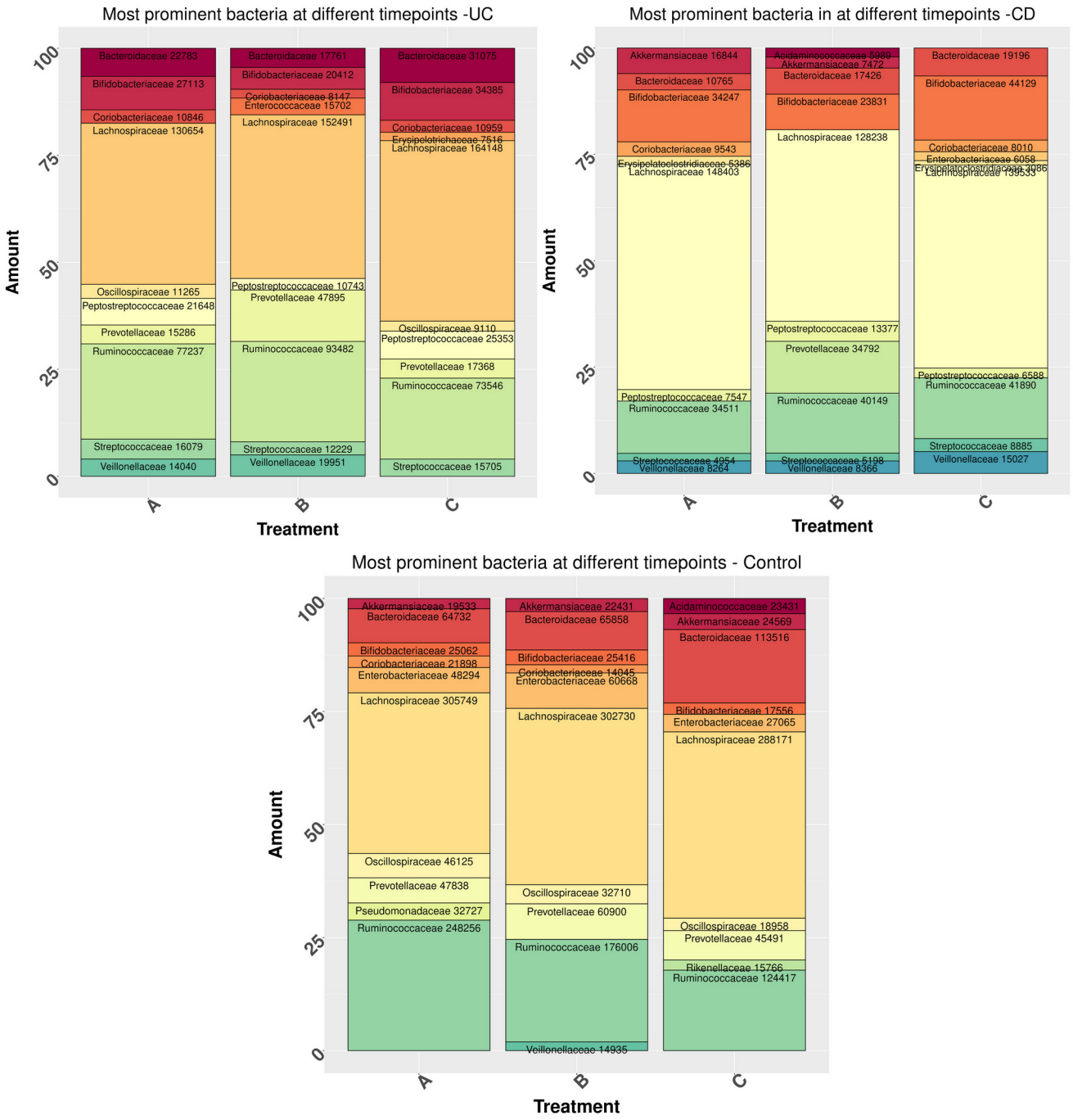


Figure 11 – Abundance tables of each sample group (UC CD and Con) at time points A B C representing the ten most abundant taxa and their counts. The Y Axis represents relative (% of total) abundance – dada2

Source: Own analysis

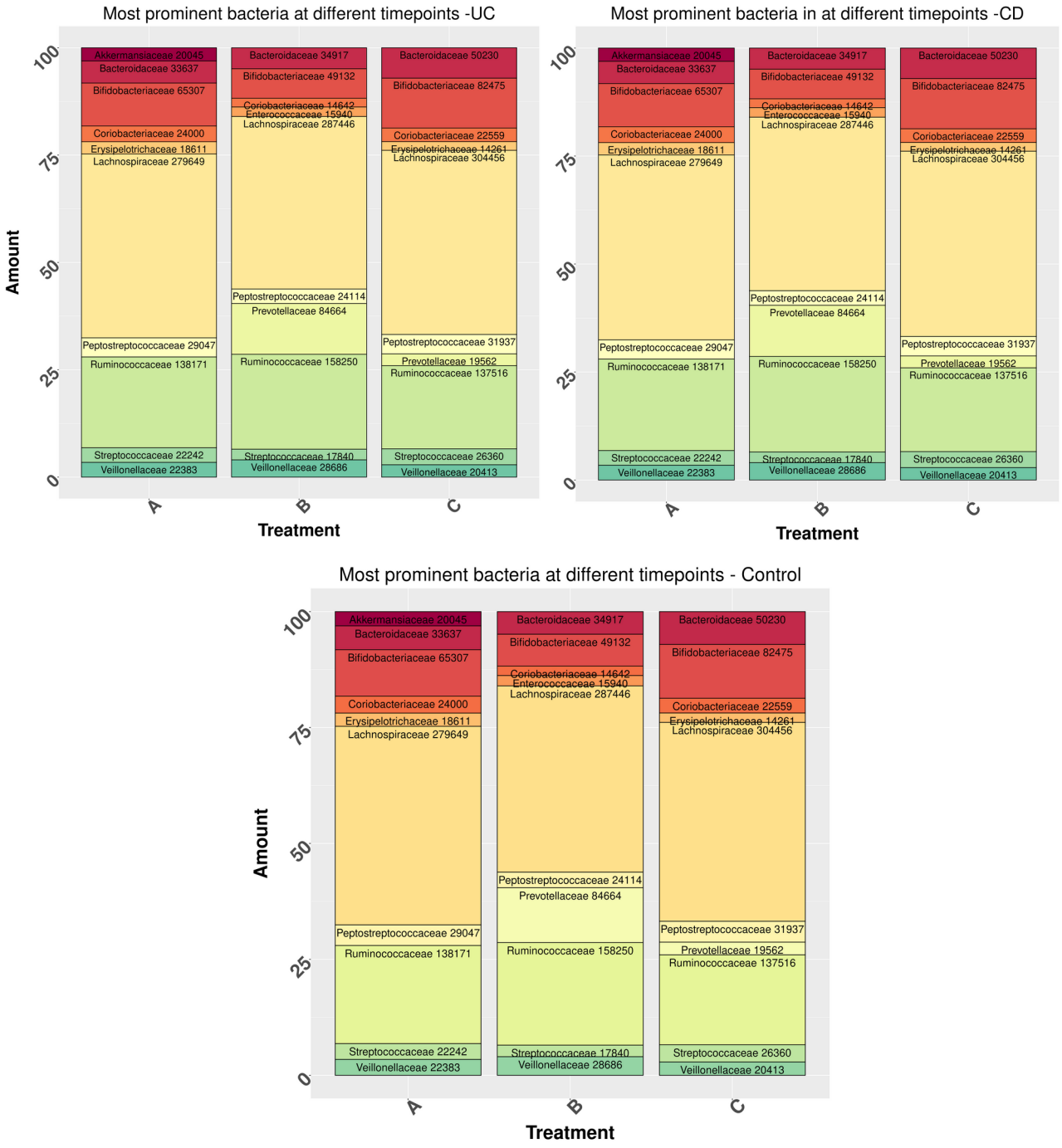


Figure 12 – Abundance tables of each sample group (UC CD and Con) at time points A B C representing the ten most abundant taxa and their counts. The Y Axis represents relative (% of total) abundance -Mothur

Source: Own analysis

IV.Discussion

At the end of pipelines both count tables were relatively similar considering parameters of mean count sums, maximum and minimum sample counts, standard deviance in sample counts (Table 3). *Mothur* however inferred more different taxa. One possible explanation for this difference is that *dada2*'s learnErrors machine learning algorithm for inferring ASVs identified more unidentical sequences to be sequencing errors, while *Mothur* assumed them to be taxonomically different due to their difference percentage. A way to further examine this would be to run *dada2* again with the OMEGA_A option (error sensitivity parameter) set lower, or run *Mothur* again but this time clustering OTU's with a higher identity cutoff (current is 99%). Many of the differences were unclassified taxa, that were higher than family level, for example Enterobacterales_unclassified, unidentified on a lower level, mostly coming from the *Mothur* analysis. On the mock sample, both pipelines performed perfectly (identifying 20 out of 20 species). Probably a better judge of performance would be a larger, more complex Mock sample.

The raw data of Shannon Alpha diversities was quite similar (Figure 4) with a few exceptions. In light of the difference of taxa, this is a strange finding, as more taxa by default should create a higher Alpha diversity. The likely reason is that the previously mentioned unclassified taxa in *Mothur* were usually not the most prevalent ones, not present in many samples, so they couldn't influence Alpha diversity as much. When comparing Alpha diversities between groups, the following differences are statistically significant:

After the bowel preparation, Alpha diversity in the CD group was lower compared to UC (p=0.0070 – *dada2*,p=0.0138 –*Mothur*) and Con (p=0.0284 -*dada2*, p=0.0215 – *Mothur*) patients at time point B. This difference continued in time point C CD was lower than control (p = 0.0215 – *dada2* p=0.0370 – *Mothur*) and UC (p = 0.0001 – *dada2*, p = 0.0005 – *Mothur*). UC was also higher than the control (p=0.0272 *Mothur*) group at time point C moreover, Alpha diversity of UC patients increased significantly from time point A to C (p=0.0441 – *Mothur*), although these findings only appear when using the *Mothur* pipeline. To summarize this, both pipelines agree that the Alpha divesities of CD patients are diminishing, compared to the other two, and Mothur also confirms that UC on the other hand is increasing. Although the climbing tendency is visible on boxplots (Figure 5, 6, 7, 8) yielded by both pipelines, only

Mothur confirms that UC is increasing over time, with the significant UC A - UC C difference. The decrease of CD continuing from B to the C time point is possibly due to the disbyosis starting after the administration of laxative, failing to regenerate. The reason for UC's upward trajectory is hard to pinpoint, but the readiest explanation is that the group of patients are regenerating due to the treatment received during the four weeks (A to C).

Beta distance PCoA analyses did not prove convincing enough due to the eigenvalues explaining very little of the diversity variance, (*dada2:* Axis 1: 10.73 % , Axis 2: 9.62 %, total: 20.35 % , *Mothur:* Axis 1: 10.98% , Axis 2: 9.44% , total: 20.44% , Figure 9, 10) although a tendency of separation between samples of ill and healthy patients was observed (Figure 10). When judging these plots, a significance cutoff line (such as 0.05 *p* value) is hard to draw, rather it should be understood, that the most determining two principal coordinates explain only 20% of the total variance in species composition, which is relatively low.

Comparisons of the mean Beta distances yielded the results in Table 7. The Beta distance between the control group's A and B time points and the CD group's A and B time points were significantly different, the higher being the distances in the control group (p= 0.0378 – *dada2*, p = 0.0352 – *Mothur*). In the case of A to C distances the UC group's was higher than the CD's (p=0.0273 – *dada2,* p=0.0138 *Mothur*). The first difference could suggest, that in the control group, the microbiome was less accustomed to diarrheic effects (caused by the laxatives), so more taxonomic changes occurred than in the group of CD, where the patients are suffering the everyday effects of IBDs. The second finding confirms that UC changed not only in raw diversity metrics (Alpha diversity) but also in taxonomical makeup, during the four week period. This again could be attributed to the medical treatment, or other unknown factors. *Mothur* has also found that the control group's A-B distances are higher than UC too, further confirming the accustomization theory.

In general I have noticed that *Mothur's* count table produced more significant results in most aspects of the analysis, which can be attributed to the higher number of different taxa, producing more variation, although this may not be the exclusive reason. Relative abundance analyses provided results both consistent and conflicting with literature, finding differences in Streptococcaceae, Clostridiaceae and Enterobacteriaceae to mention the most notable.

Streptococcaceae showed growth from the A to B time point in CD compared to Con. The taxon has been associated associatied with IBDs (Hertogh et al. 2008).

Clostridiaceae was generally more abundant in control patients, which is contrary to what literature suggests. Clostridiaceae is one of the taxa that are most commoly associated with IBDs (Azimi et al. 2018, Wright et al. 2017) .The increases were all around the samples, not only in individual Patients. The group Enterobacteriaceae, that musters the infamous of *Escherichia coli* in its ranks along with many other detrimental taxa that are associated with IBD (Vester-Andersen et al. 2019, Wright et al. 2017,Zuo and Ng 2018,) was found to be higher in the control samples than in the CD group at time point C. This also seems to contradict literature, but it might be due to outbreaks not associated with the IBDs, although the increases were fairly even. Since Enterobacteriaceae contains many opportunistic pathogens, this could also be attributed to outbreaks unrelated to the illnesses or the treatment. Pseudomonaceae has been observed to have an abundance change in two individuals (37 and 41). Patient 41's decrease in Alpha diversity due to one taxa (Pseuodomonaceae) becoming extremely dominant is visible on Figure 4.

To compare the two pipelines from the user's perspective, I will present a short discussion about each, then compare their strengths and weaknesses. In *dada2*, everything is handled inside *RStudio*, all of the data is also saved in the project environment, instead of the folder by default, although there is nothing preventing the user from making backups manually. A great advantage of *RStudio* integration is that the user already has the knowledge to configure the pipeline's formats to his or her dataset's unique format.

The bioinformatics tool *Mothur* is not integrated in *R*, rather it's a command line tool that can be used in linux's or macOS's bash terminal or Windows' shell, since it's available on all three platforms. Once it's installed it's interface can be invoked by simply typing *Mothur* in the command line. As it can be concluded from the pipeline summary, *Mothur* does not rely as much on plotting results instead it uses text based summaries. Since it's not integrated in *R* it is also not dependent on any packages, but it relies on the following external tools: (*blast, vsearch, uchime, prefetch* and *fasterq-dump*), which are included in the code.

In comparison I think *dada2*'s advantages are *RStudio* integration, consequently better visualization options during the workflow, and ASV level analysis. It lacks however an automatic backup system, and is less 'stand-alone' than *Mothur*, as the latter can be ran from

any operation system's command line. *Mothur*'s advantages are that it was around twice as fast in every step, and automatic backup system was really intuitive.

To conclude the discussion, while significant differences were observed in all areas of this study, limitations of our methods apply: further investigation with more participants, and using more 16S variable regions is advised, since V4 is only considered a valid indicator of taxa origin for determining family level identity.

V.Summary:

In the introduction of my thesis, I describe the general properties of the human microbiome and talk about its significance in our well-being, presenting the characteristics of the two discussed IBDs; Ulcerative colitis and Crohn's disease, mentioning similarities and differences. Next, I show key elements of the microbiome investigation process, such as the 16S region, hypervariable regions, particularly V4, barcode sequences and dataset. Differences between Amplicon and Shotgun sequencing are also highlighted. In the first part of the Methods section the microbiome analysis pipeline is detailed, describing each crucial step. In the second part, I detail the post-classification steps. Examination of 43 people's samples out of whom 19 were healthy (control), 13 treated with Ulcerative Colitis, 9 with Crohn's disease, resulted in 123 samples belonging to time points A B and C, containing abundance data in a count table created by *Mothur* and *dada2*, two different 16S rRNA sequence analysis pipelines. The mock sample I used contained 20 known strains altogether. To evaluate the classification the user can compare the inferred sequence to the expected makeup of the Mock community. Both *dada2* and *Mothur* inferred 20 out of 20 sequences. The data gathered was investigated with *R* using the *RStudio* IDE. The following analyses were performed: Alpha diversity, Beta diversity and relative abundance comparison. Alpha diversity yielded the following significant results: The CD group was lower compared to UC and Con patients at time point B. This difference continued in time point C: CD was lower than control and UC. UC was also higher than the control group at time point C moreover, Alpha diversity of UC patients increased significantly from time point A to C, although these findings only appear when using the *Mothur* pipeline. Beta distance PCoA analyses did not prove convincing enough due to the eigenvalues explaining very little of the diversity variance, (Figure 9, 10) although a tendency of separation between samples of ill and healthy patients was observed (Figure 10). The Beta distance between the control group's A and B time points and the CD group's A and B time points were significantly different, the higher being the distances in the control group. In the case of A to C distances the UC group's was higher than the CD's. In the discussion section, I aim to put these results into context, and emphasize the conclusions that can be drawn, also noting shortcomings of the results, and limitations of our methods.

STATEMENT

Name: Asbóth András
Neptun ID: D5I2GO

ELTE Faculty of Science: Biology MSc

specialization: Molecular Immune- and Microbiology

Title of diploma work: A comparative look at different microbiome analysis pipelines

As the author of the diploma work I declare, with disciplinary responsibility that my thesis is my own intellectual product and the result of my own work. Furthermore I declare that I have consistently applied the standard rules of references and citations.

I acknowledge that the following cases are considered plagiarism:
– using a literal quotation without quotation mark and adding citation;
– referencing content without citing the source;
– representing another person's published thoughts as my own thoughts.

Furthermore, I declare that the printed and electronical versions of the submitted diploma work are textually and contextually identical.

Budapest, 2020.05.15

Signature of Student

VI.References

Acinas, Silvia G., Ramahi Sarma-Rupavtarm, Vanja Klepac-Ceraj, and Martin F. Polz. 2005. "PCR-Induced Sequence Artifacts and Bias: Insights from Comparison of Two 16S RRNA Clone Libraries Constructed from the Same Sample." *Applied and Environmental Microbiology* 71 (12): 8966–69.

Appanna, Vasu D. 2018. "The Microbiome: Genesis and Functions." In *Human Microbes - The Power Within: Health, Healing and Beyond*, edited by Vasu D. Appanna, 37–79. Singapore: Springer.

Azimi, Taher, Mohammad Javad Nasiri, Alireza Salimi Chirani, Ramin Pouriran, and Hossein Dabiri. 2018. "The Role of Bacteria in the Inflammatory Bowel Disease Development: A Narrative Review." *APMIS: Acta Pathologica, Microbiologica, et Immunologica Scandinavica* 126 (4): 275–83.

Bäckhed, Fredrik, Ruth E. Ley, Justin L. Sonnenburg, Daniel A. Peterson, and Jeffrey I. Gordon. 2005. "Host-Bacterial Mutualism in the Human Intestine." *Science (New York, N.Y.)* 307 (5717): 1915–20.

Bukin, Yu S., Yu P. Galachyants, I. V. Morozov, S. V. Bukin, A. S. Zakharenko, and T. I. Zemskaya. 2019. "The Effect of 16S RRNA Region Choice on Bacterial Community Metabarcoding Results." *Scientific Data* 6 (1): 190007.

Callahan, Benjamin J., Paul J. McMurdie, and Susan P. Holmes. 2017. "Exact Sequence Variants Should Replace Operational Taxonomic Units in Marker-Gene Data Analysis." *The ISME Journal* 11 (12): 2639–43.

Chakravorty, Soumitesh, Danica Helb, Michele Burday, Nancy Connell, and David Alland. 2007. "A Detailed Analysis of 16S Ribosomal RNA Gene Segments for the Diagnosis of Pathogenic Bacteria." *Journal of Microbiological Methods* 69 (2): 330–39.

Clarridge, Jill E. 2004. "Impact of 16S RRNA Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases." *Clinical Microbiology Reviews* 17 (4): 840–62, table of contents.

Collins, Paul, and Jonathan Rhodes. 2006. "Ulcerative Colitis: Diagnosis and Management." *BMJ : British Medical Journal* 333 (7563): 340–43.

Dahlhamer, James M., Emily P. Zammitti, Brian W. Ward, Anne G. Wheaton, and Janet B. Croft. 2016. "Prevalence of Inflammatory Bowel Disease Among Adults Aged ≥18 Years - United States, 2015." *MMWR. Morbidity and Mortality Weekly Report* 65 (42): 1166–69.

Dekaboruah, Elakshi, Mangesh Vasant Suryavanshi, Dixita Chettri, and Anil Kumar Verma. 2020. "Human Microbiome: An Academic Update on Human Body Site Specific Surveillance and Its Possible Role." *Archives of Microbiology* 202 (8): 2147–67.

Dillies, Marie-Agnès, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Céline Keime, et al. 2013. "A Comprehensive Evaluation of Normalization Methods for Illumina High-Throughput RNA Sequencing Data Analysis." *Briefings in Bioinformatics* 14 (6): 671–83.

Ewing, Brent, LaDeana Hillier, Michael C. Wendl, and Phil Green. 1998. "Base-Calling of Automated Sequencer Traces UsingPhred. I. Accuracy Assessment." *Genome Research* 8 (3): 175–85.

Fuks, Garold, Michael Elgart, Amnon Amir, Amit Zeisel, Peter J. Turnbaugh, Yoav Soen, and Noam Shental. 2018. “Combining 16S RRNA Gene Variable Regions Enables High-Resolution Microbial Community Profiling.” *Microbiome* 6 (1): 17.

Gajendran, Mahesh, Priyadarshini Loganathan, Anthony P. Catinella, and Jana G. Hashash. 2018. “A Comprehensive Review and Update on Crohn’s Disease.” *Disease-a-Month: DM* 64 (2): 20–57.

García-López, Rodrigo, Fernanda Cornejo-Granados, Alonso A. Lopez-Zavala, Filiberto Sánchez-López, Andrés Cota-Huízar, Rogerio R. Sotelo-Mundo, Abraham Guerrero, Alfredo Mendoza-Vargas, Bruno Gómez-Gil, and Adrian Ochoa-Leyva. 2020. “Doing More with Less: A Comparison of 16S Hypervariable Regions in Search of Defining the Shrimp Microbiota.” *Microorganisms* 8 (1).

Gevers, Dirk, Subra Kugathasan, Lee A. Denson, Yoshiki Vázquez-Baeza, Will Van Treuren, Boyu Ren, Emma Schwager, et al. 2014. “The Treatment-Naive Microbiome in New-Onset Crohn’s Disease.” *Cell Host & Microbe* 15 (3): 382–92.

Ghosh, Srikanta, and Simpson Joseph. 2005. “Nonbridging Phosphate Oxygens in 16S RRNA Important for 30S Subunit Assembly and Association with the 50S Ribosomal Subunit.” *RNA (New York, N.Y.)* 11 (5): 657–67.

Guiso, Nicole. 2015. “Chapter 85. Bordetella Pertussis.” In , 1507–27.

Gupta, Vinod K., Sandip Paul, and Chitra Dutta. 2017. “Geography, Ethnicity or Subsistence-Specific Variations in Human Microbiome Composition and Diversity.” *Frontiers in Microbiology* 8.

He, Yan, Wei Wu, Hui-Min Zheng, Pan Li, Daniel McDonald, Hua-Fang Sheng, Mu-Xuan Chen, et al. 2018. “Regional Variation Limits Applications of Healthy Gut Microbiome Reference Ranges and Disease Models.” *Nature Medicine* 24 (10): 1532–35.

Hertogh, Gert De, Jeroen Aerssens, Karen P Geboes, and Karel Geboes. 2008. “Evidence for the Involvement of Infectious Agents in the Pathogenesis of Crohn’s Disease.” *World Journal of Gastroenterology : WJG* 14 (6): 845–52.

Hughes, Riley L. 2020. “A Review of the Role of the Gut Microbiome in Personalized Sports Nutrition.” *Frontiers in Nutrition* 6.

Janda, J. Michael, and Sharon L. Abbott. 2007. “16S RRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls.” *Journal of Clinical Microbiology* 45 (9): 2761–64.

Jolliffe, I. T., ed. 2002. “Introduction.” In *Principal Component Analysis*, 1–9. Springer Series in Statistics. New York, NY: Springer.

Khan, Israr, Naeem Ullah, Lajia Zha, Yanrui Bai, Ashiq Khan, Tang Zhao, Tuanjie Che, and Chunjiang Zhang. 2019. “Alteration of Gut Microbiota in Inflammatory Bowel Disease (IBD): Cause or Consequence? IBD Treatment Targeting the Gut Microbiome.” *Pathogens (Basel, Switzerland)* 8 (3).

Lemmens, B., G. De Hertogh, and X. Sagaert. 2014. “Inflammatory Bowel Diseases.” In *Pathobiology of Human Disease*, edited by Linda M. McManus and Richard N. Mitchell, 1297–1304. San Diego: Academic Press.

Lichtenstein, Gary R. 2010. “Current Research in Crohn’s Disease and Ulcerative Colitis: Highlights from the 2010 ACG Meeting.” *Gastroenterology & Hepatology* 6 (12 Suppl 17): 3–14.

Liu, Richard T. 2017. “The Microbiome as a Novel Paradigm in Studying Stress and Mental Health.” *American Psychologist* 72 (7): 655–67.

Loddo, Italia, and Claudio Romano. 2015. “Inflammatory Bowel Disease: Genetics, Epigenetics, and Pathogenesis.” *Frontiers in Immunology* 6 (November).

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *Genome Biology* 15 (12): 550.

Luo, Annie, Steven T. Leach, Romain Barres, Luke B. Hesson, Michael C. Grimm, and David Simar. 2017. “The Microbiota and Epigenetic Regulation of T Helper 17/Regulatory T Cells: In Search of a Balanced Immune System.” *Frontiers in Immunology* 8.

Manor, Ohad, Chengzhen L. Dai, Sergey A. Kornilov, Brett Smith, Nathan D. Price, Jennifer C. Lovejoy, Sean M. Gibbons, and Andrew T. Magis. 2020. “Health and Disease Markers Correlate with Gut Microbiome Composition across Thousands of People.” *Nature Communications* 11 (1): 5206.

M’Koma, Amosy E. 2013. “Inflammatory Bowel Disease: An Expanding Global Health Problem.” *Clinical Medicine Insights. Gastroenterology* 6: 33–47.

Moore, Edward R.B., Margit Mau, Angelika Arnscheidt, Erik C. Böttger, Roger A. Hutson, Matthew D. Collins, Yves Van De Peer, Rupert De Wachter, and Kenneth N. Timmis. 1996. “The Determination and Comparison of the 16S RRNA Gene Sequences of Species of the Genus Pseudomonas (Sensu Stricto and Estimation of the Natural Intrageneric Relationships.” *Systematic and Applied Microbiology* 19 (4): 478–92.

Ni, Josephine, Gary D. Wu, Lindsey Albenberg, and Vesselin T. Tomov. 2017. “Gut Microbiota and IBD: Causation or Correlation?” *Nature Reviews Gastroenterology & Hepatology* 14 (10): 573–84.

Oksanen, Jari, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, et al. 2020. *Vegan: Community Ecology Package*. https://CRAN.R-project.org/package=vegan.

Petti, C. A., C. R. Polage, and P. Schreckenberger. 2005. “The Role of 16S RRNA Gene Sequencing in Identification of Microorganisms Misidentified by Conventional Methods.” *Journal of Clinical Microbiology* 43 (12): 6123–25.

Pollock, Jolinda, Laura Glendinning, Trong Wisedchanwet, and Mick Watson. 2018. “The Madness of Microbiome: Attempting To Find Consensus ‘Best Practice’ for 16S Microbiome Studies.” *Applied and Environmental Microbiology* 84 (7).

Porter, Teresita M., and Mehrdad Hajibabaei. 2018. “Scaling up: A Guide to High-Throughput Genomic Approaches for Biodiversity Analysis.” *Molecular Ecology* 27 (2): 313–38.

Prodan, Andrei, Valentina Tremaroli, Harald Brolin, Aeilko H. Zwinderman, Max Nieuwdorp, and Evgeni Levin. 2020. “Comparing Bioinformatic Pipelines for Microbial 16S RRNA Amplicon Sequencing.” *PloS One* 15 (1): e0227434.

“Release Version 1.45.2 · Mothur/Mothur.” n.d. GitHub. Accessed April 22, 2021. /mothur/mothur/releases/tag/v1.45.2.

Rhoads, Anthony, and Kin Fai Au. 2015. “PacBio Sequencing and Its Applications.” *Genomics, Proteomics & Bioinformatics*, SI: Metagenomics of Marine Environments, 13 (5): 278–89.

Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2010. “EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data.” *Bioinformatics* 26 (1): 139–40.

Robinson, Mark D., and Alicia Oshlack. 2010. “A Scaling Normalization Method for Differential Expression Analysis of RNA-Seq Data.” *Genome Biology* 11 (3): R25.

Scaldaferri, Franco, Viviana Gerardi, Loris Riccardo Lopetuso, Fabio Del Zompo, Francesca Mangiola, Ivo Boškoski, Giovanni Bruno, et al. 2013. “Gut Microbial Flora, Prebiotics, and Probiotics in IBD: Their Current Usage and Utility.” *BioMed Research International* 2013: 435268.

Schirmer, Melanie, Sanne P. Smeekens, Hera Vlamakis, Martin Jaeger, Marije Oosting, Eric A. Franzosa, Rob ter Horst, et al. 2016. “Linking the Human Gut Microbiome to Inflammatory Cytokine Production Capacity.” *Cell* 167 (4): 1125-1136.e8.

Schloss, Patrick D., Sarah L. Westcott, Thomas Ryabin, Justine R. Hall, Martin Hartmann, Emily B. Hollister, Ryan A. Lesniewski, et al. 2009. “Introducing Mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities.” *Applied and Environmental Microbiology* 75 (23): 7537–41.

Thia, Kelvin T., Edward V. Loftus, William J. Sandborn, and Suk-Kyun Yang. 2008. “An Update on the Epidemiology of Inflammatory Bowel Disease in Asia.” *The American Journal of Gastroenterology* 103 (12): 3167–82.

Tyson, John R., Nigel J. O’Neil, Miten Jain, Hugh E. Olsen, Philip Hieter, and Terrance P. Snutch. 2018. “MinION-Based Long-Read Sequencing and Assembly Extends the Caenorhabditis Elegans Reference Genome.” *Genome Research* 28 (2): 266–74.

Ungaro, Ryan, Saurabh Mehandru, Patrick B Allen, Laurent Peyrin-Biroulet, and Jean-Frédéric Colombel. 2017. “Ulcerative Colitis.” *The Lancet* 389 (10080): 1756–70.

Varet, Hugo, Loraine Brillet-Guéguen, Jean-Yves Coppée, and Marie-Agnès Dillies. 2016. “SARTools: A DESeq2- and EdgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data.” *PloS One* 11 (6): e0157022.

Vester-Andersen, M. K., H. C. Mirsepasi-Lauridsen, M. V. Prosberg, C. O. Mortensen, C. Träger, K. Skovsen, T. Thorkilgaard, et al. 2019. “Increased Abundance of Proteobacteria in Aggressive Crohn’s Disease Seven Years after Diagnosis.” *Scientific Reports* 9 (1): 13473.

Voreades, Noah, Anne Kozil, and Tiffany L. Weir. 2014. “Diet and the Development of the Human Intestinal Microbiome.” *Frontiers in Microbiology* 5.

Wang, Qiong, George M. Garrity, James M. Tiedje, and James R. Cole. 2007. “Naive Bayesian Classifier for Rapid Assignment of RRNA Sequences into the New Bacterial Taxonomy.” *Applied and Environmental Microbiology* 73 (16): 5261–67.

Wei, Bo, Tiffany Huang, Harnisha Dalwadi, Christopher L. Sutton, David Bruckner, and Jonathan Braun. 2002. “Pseudomonas Fluorescens Encodes the Crohn’s Disease-Associated I2 Sequence and T-Cell Superantigen.” *Infection and Immunity* 70 (12): 6567–75.

Wright, Emily K., Michael A. Kamm, Josef Wagner, Shu-Mei Teo, Peter De Cruz, Amy L. Hamilton, Kathryn J. Ritchie, Michael Inouye, and Carl D. Kirkwood. 2017. “Microbial Factors Associated with Postoperative Crohn’s Disease Recurrence.” *Journal of Crohn’s and Colitis* 11 (2): 191–203.

Yang, Bo, Yong Wang, and Pei-Yuan Qian. 2016. “Sensitivity and Correlation of Hypervariable Regions in 16S RRNA Genes in Phylogenetic Analysis.” *BMC Bioinformatics* 17 (March).

Yu, Yangyang R., and J. Ruben Rodriguez. 2017. “Clinical Presentation of Crohn’s, Ulcerative Colitis, and Indeterminate Colitis: Symptoms, Extraintestinal Manifestations, and Disease Phenotypes.” *Seminars in Pediatric Surgery* 26 (6): 349–55.

Zuo, Tao, and Siew C. Ng. 2018. “The Gut Microbiota in the Pathogenesis and Therapeutics of Inflammatory Bowel Disease.” *Frontiers in Microbiology* 9 (September).

Acknowledgements

I would like to thank my supervisor Eszter Ari, for introducing me to the field of bioinformatics, more specifically to the analysis of 16S rRNA data, and for the support she lent to my work. I would also like to thank Mariann Rutka and Ferenc Molnár who collaborated by providing the samples and metadata for this thesis.