UNIVERSITAT DE
BARCELONA

The human gut microbiome and its influence in mental health

Eduardo Herreros Valenzuela
Máster Universitario en Bioinformática y Bioestadística
TFM – Bioinformática y Bioestadística Área 3

Andreu Paytuví Gallart
Ferran Prados Carrasco

08 de enero de 2020

FICHA DEL TRABAJO FINAL

Título del trabajo:	*The human gut microbiome and its influence in mental health*
Nombre del autor:	*Eduardo Herreros Valenzuela*
Nombre del consultor/a:	*Andreu Paytuví Gallart*
Nombre del PRA:	*Ferran Prados Carrascos*
Fecha de entrega (mm/aaaa):	01/2020
Titulación:	*Máster Universitario en Bioinformática y Bioestadística*
Área del Trabajo Final:	*TFM – Bioinformática y Bioestadística Área 3*
Idioma del trabajo:	*Inglés*
Palabras clave	*Gut microbiome, Mental health, Gut-Brain axis, Machine Learning.*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

Se estima que problemas en la salud mental afectan al 25% de la población, lo que la convierte en la principal causa de enfermedades a nivel mundial. Una microbiota normal está relacionada con estados saludables, sin embargo, los cambios en su composición (llamada disbiosis) están relacionados con enfermedades. En este proyecto, exploramos la influencia de la microbiota intestinal en la aparición de estados mentales no saludables utilizando enfoques de machine learning, clasificación de enterotipos y análisis estadísticos univariados y multivariados.

Entre las características demográficas encontramos diferencias entre los estados de la enfermedad mental en la etnia ($p = 0.04$), sexo ($p = 0.004$), enfermedad del intestino irritable ($p < 0.001$), etc. Encontramos niveles más bajos de *Firmicutes* y niveles más altos de *Bacteroidetes* y una proporción más baja de Firmicutes/Bacteroidetes en el intestino de las personas con problemas mentales. Las personas con enfermedades mentales tienen un índice de diversidad alfa más bajo en su intestino en comparación con las personas sanas ($p = 0.002$). El análisis de diversidad beta presentó diferentes centroides con respecto a los estados mentales, por la prueba PERMANOVA ($p = 0.032$). El mejor predictor de machine learning fue Random Forest con una precisión de 0.62. Sin embargo, probablemente al mezclar los diferentes trastornos mentales que tienen un fondo biológico diferente, creamos ruido haciendo que los resultados de predicción de los algoritmos de machine learning tengan rendimientos bajo los esperados.

En conclusión, una concentración de *Firmicutes* más altos y de *Bacteroidetes*

más bajos podrían actuar como un factor de riesgo en la aparición de un trastorno mental

Abstract (in English, 250 words or less):

Mental health problems affect 25% of the population, making it the leading cause of disability globally. Normal microbiota is related with healthy states, however changes in its composition (called dysbiosis) is linked with non-healthy pathologies. In this project we explored the influence of the gut microbiota on the occurrence of non-healthy mental states using machine learning approaches, enterotype classifications and univariate and multivariate statistical analyses.

Among the demographic characteristics we found differences between the mental illness states in the ethnicity ($p = 0.04$), sex ($p = 0.004$), irritable bowel disease ($p < 0.001$), etc. We found lower levels of *Firmicutes* and higher levels of *Bacteroidetes* and a lower Firmicutes/Bacteroidetes ratio in the gut of people with mental issues. People with mental illness has a lower alpha diversity index in their gut in comparison with healthy people ($p = 0.002$). The beta diversity analysis presented different centroids regarding the mental states statistically measured by the PERMANOVA test ($p = 0.032$). The best machine learning predictor was Random Forest with an accuracy of 0.62. However, because of we mixed the different mental disorders with a different biological background, probably creating noise, the prediction results of the machine learning algorithms do not have better performances.

In conclusion, more efforts are necessary in the use of machine learning algorithms with microbiome information, because of the potential that these methods have in the classification and/or prediction of certain pathologies. Also, higher *Firmicutes* and lower *Bacteroidetes* could be risk factor in the occurrence of a mental illness.

Table of contents

1. Introduction 1
1.1 Work context and justification 1
1.2 Objectives 2
1.3 Method followed 3
1.4 Working program 3
1.5 Brief summary of products obtained 3
1.6 Brief description of the other chapters of the thesis 3
2. Methods 4
2.1 Data preparation, manipulation and normalization. 4
2.2 Statistical analysis 4
2.3 Enterotypes analysis 5
2.4 Differential abundance analysis 5
2.5 Machine learning algorithms 5
3. Results 6
3.1 Baseline characteristics of the study population. 6
3.2 Gut microbiota composition in the study population. 7
3.3 Alpha and Beta-diversity of the gut microbiota in mental illness population. 9
3.4 Enterotype of the gut microbiota of the study population 11
3.5 Differential abundance analysis of the gut microbiota 12
3.6 Mental illness prediction by machine learning algorithms. 14
4. Discussion 15
5. Conclusions 17
6. Glossary 18
7. References 19

List of figures

Figure 1: Phylum composition of bacteria from gut samples of people with or without mental illness.
Figure 2: Top 5 Phylum order of gut samples of people with or without mental illness.
Figure 3: Top 5 Family order of gut samples of people with or without mental illness.
Figure 4: Top 5 Genus order of gut samples of people with or without mental illness.
Figure 5: Boxplots of diversity indices for the gut microbiome, Chao1 and Shannon indices.
Figure 6: Ordination (PCoA) generated by using the Bray–Curtis dissimilarity metric for the two mental illness states. Samples are colored according to mental illness state.
Figure 7: Optimal clusters number.
Figure 8: Between-class analysis (BCA) of the enterotypes.
Figure 9: Phylum and family order representation of the DESeq2 results.
Figure 10: Predictive accuracy comparison of the machine learning models.
Figure 11: Machine learning models prediction accuracy using the training and test dataset

1. Introduction

1.1 Work context and justification

General description

Microbiome is defined as the genome of a microbiota. On the other side, microbiota is the wide collection of microbes which conforms a microbial community (viruses, bacteria, yeasts, etc.) in one specific environment. Bacterial cells in the human body account for 1-3% of total body weight and are at least equal in number to human cells. Several recent research studies have been focused to understand how the different bacterial communities in the body (e.g. gut, respiratory tract, skin, and vaginal microbiota) predispose to health and disease. For instance, the genus Lactobacillus is often found in the vagina while Bacteroidetes and Prevotella in the gut. Normal microbiota is related with healthy states, however changes in its composition (called dysbiosis) is linked with non-healthy pathologies.

Thanks to scientific advances, nowadays we are able to analyze microbial communities through Next Generation Sequencing (NGS) techniques. The most used sequencing methods are Whole Genome Sequencing (WGS) and rDNA sequencing. NGS is particularly used in projects as “The Human Microbiome Project” (HMP). The aim of HMP is to map the whole human genome. Moreover, WGS opened the opportunity to a wide spectrum of studies that can go from decipher what are the effects of one mutation can provoke in each organism, to evolutive studies and how the DNA is changing. 16S rDNA is widely used in the field of microbial ecology. The 16S rRNA gene is a DNA region highly conserved among bacteria with hypervariable regions within it. With the use of specific primers attached to the conserved region of 16S rRNA gene we can amplify the segment of interest and to identify, at a phylogenetic level, which genus or species of bacteria dominates the different microbiotas. One issue about 16S rDNA sequencing is that only recognize bacterial species and not viruses or yeasts.

Bioinformatic advances let us explore and interpret all the data recollected from the sequencing techniques. Multidimensional cluster analyses and principal component analyses (PCA) are used to group the different kind of microbiotas; this is called enterotyping. This can be useful in the identification of dysbiotic states or microbial patterns that can be correlated to certain pathology.

Master’s Thesis justification

Mental health problems affect 25% of the population, making it the leading cause of disability globally (1,2). United Nation’s (UN) states that good mental health is fundamental to individual’s well-being and overall health (3,4). During the past years, mental health has predominantly understood as related to the brain affections, however nowadays the efforts are focused on revealing how systemic and environmental changes could be affecting in this field (5, 6). Recent studies confirm evident links between mental disorders (e. g. Alzheimer) and gut microbiota, suggesting that mental health is not only related to the brain (7, 8).

Intestinal microbiota is one of the richest ecosystems in nature. Usually, the common belief was that the gut microbiome was only acting in the digestive process, however thanks to molecular and sequencing procedures, worldwide scientists have been disclosing the role of microbial interaction with countless physiological processes (9, 10). Currently, scientists believe about the existence of a microbiome-gut-brain axis (8).

Intestinal microbiota integrates part of the intestinal barrier, which controls the transport of antigens. Thus, the gut microbiota can be influencing the intestinal barrier permeability guaranteeing that only some molecules will go through it to the blood vessels (11, 12). When a gut dysbiotic event happens, intestinal permeability rises, causing an increase of inflammatory mediators and effector cells disrupting gut barrier (13).

Machine learning (ML) methods have been widely applied in the field of microbial ecology. ML could help dealing with group classification problems regarding microbiota and explore the interaction between microorganisms and the environment that they conform. Recently, studies have begun to explore the power of ML to use microbiome patterns to predict host characteristics (14, 15). However, the content-knowledge required to implement these methods is high, presenting a barrier to data scientists looking to get started in microbiome analysis and prediction.

Our main goal with this project was explore the influence of the gut microbiota on the occurrence of non-healthy mental states using machine learning approaches, enterotype classifications and univariate and multivariate statistical analyses.

1.2 Objectives

Main Objective

To evaluate whether gut microbiota differs along the different healthy and non-healthy mental states with data from the American Gut Project.

Specific Objectives

For accomplish the main objective we defined four specific objectives:

1) To study the gut microbial diversity (alpha and beta) among the mental (healthy or non-healthy) states.
2) To analyze the gut microbial dissimilarity through enterotype clusters formation in the healthy and non-healthy mental states.
3) To perform gut microbial differential analyses, univariate and multivariate analyses in the study cases.
4) To apply machine learning algorithms as predictors of differential features of the non-healthy mental state.

1.3 Method followed

Gut microbiome 16S rDNA data analysis using R for the data preparation, data mining, and machine learning predictive models' development. Microbiome description; Alpha and beta diversity, Enterotype and DESeq2 analyses. The methods are extensively detailed in the section 2.1.

1.4 Working program

Activities	September	October				November				December				January			
	4th	1st	2nd	3rd	4th	1st	2nd	3rd	4rd	1st	2nd	3rd	4rd	1st	2nd	3rd	4rd
Research Project																	
PEC0	D1																
PEC1		D2															
PEC2				D3													
Data preparation				M1													
Alpha and Beta diversity						M2											
Enterotypes							M3										
PEC3								D4									
Univariate analyses								M4									
Multivariate analyses									M5								
Machine Learning analyses										M6							
PEC4												D5					
Thesis writing												M7					
PEC5a															D6		
PEC5b																D7	

Milestones

M1: 15/10 - 29/10. Data Preparation: transformation, normalization, taxonomy annotation, data frame creation

M2: 30/10 - 08/11. Alpha and Beta diversity analysis: Shannon and Chao1 indeces, UniFrac measurement, PCoA representation

M3: 09/11 - 18/11. Enterotypes analysis: Jensen-Shannon divergence matrix, PAM clustering, PCoA representation

M4: 19/11 - 26/11. Univariate analyses: Mann Whitney U Test, Chi-square Test

M5:27/11 - 04/12. Multivariate analyses: DESeq2, PERMANOVA

M6: 05/12 - 16/12. Machine learning analyses: LDA, SVM, *k*-NN, Random Forest, Gradient Boosting, Neural Networks

M7: 17/12 - 08/01. Thesis writing: Introduction, methodology, results, discussion, conclusion

Derivables

D1: 18/09 - 30/09. PEC0: Work contents definition

D2: 01/10 - 14/10. PEC1: Work Program

D3: 15/10 - 18/11. PEC2: Work development 1. Data preparation and diversity representation

D4: 19/11 - 16/12. PEC3: Statistical and Machine Learning analyses

D5: 17/12 - 08/01. PEC4: Thesis writing and submission

D6: 09/01 - 12/01. PEC5a: Presentation preparation

D7: 15/01 - 22/01. PEC5b: Public Thesis defense

1.5 Brief summary of products obtained

The products are this memory, seven machine learning models and a GitHub repository with the R code that was utilized for the entire master thesis (https://github.com/EduardoHerreros/Master-Thesis).

1.6 Brief description of the other chapters of the thesis

In the other chapters you will find:

Methods: detailed method followed in the preparation of this thesis.
Results: All the results that we considered relevant to achieve the objectives of this thesis.
Discussion: Contrast of our results with previous scientific papers regarding the same topics.

2. Methods

2.1 Data preparation, manipulation and normalization.

Taxonomy matrices were supplied by Andreu Paytuví which were collected from the American Gut project. A sample data set was created from multiples JSON archives. Columns related with mental health problems were chosen, also, those ones that can acts as possible confounding factors, as sex, alcohol consumption, antibiotic consumption BMI, etc. Missing data was replaced by NAs. To have a complete raw counts OTU table is critical because machine learning methods have difficulty with missing features. This data frame creation was done using R. Only people with follow criteria were added to the analysis:

- Older than 18 years old
- Height between 1.40 and 2.10 meters
- Weight between 45 and 200 kilograms
- Maximum BMI of 50

Mental illness variable was created using the information of the variable: Attention deficit disease (ADD, ADHD), Alzheimer's disease, Autist syndrome (ASD), Epilepsy or seizure disorder, Anorexia nervosa, Bipolar disorder, Bulimia nervosa, Depression, Post-traumatic syndrome disease (PTSD), Schizophrenia and Substance abuse. With the presence of any of these variables is considered "yes" for Mental illness variable. Finally, data was weighted choosing randomly the same amount of "yes" and "no" in Mental illness variable.
To ensure compatibility of the data across the samples, data was normalized using relative abundance for the abundance plots and rarefy to its depth in the alpha and beta diversity analyses.

2.2 Statistical analysis

Statistical analysis of demographic data and mental illness outcomes were performed using R version 3.6.1. Also, it was used for data mining, multivariate analysis, enterotypes analysis and machine learning predictive models. The normality of the data was tested using the Lillie test (Kolmogorov-Smirnov normality test) from nortest library. Continuous variables were presented as median (25th, 75th percentile) and analyzed using Mann Whitney U test. Categorical variables were presented as numbers or percentages (%) an analyzed using chi-square. An OTU count table, taxonomy classification table, related clinical and demographic data and the OTU table were analyzed as a 'phyloseq' object, allowing a unified and interactive analysis approach.
OTU tables were rarefied to even depth. To assess alpha-diversity, richness (Chao1 index) and microbial-diversity (Shannon index) were computed at OTU level. Wilcoxon-rank test was performed to compare control and mental illness diversity among them. Beta-diversity using Bray Curtis distance (non-phylogenetic) was used as input for ordination analysis using principal component analysis (PCoA). In order to study the centroids differences, distance matrices analysis was performed, using adonis test (PERMANOVA)

from vegan library. Permutation test for homogeneity of multivariate dispersions was applied using the permutest.
All differences were considered statistically significant with a $p < 0.05$ with two-sided alternative hypotheses.

2.3 Enterotypes analysis

To gain overall insight into the microbial compositional structure of both; controls and mental illness cases, we applied a previously described clustering approach (16). Samples were clustered based on relative abundances using JSD distance and the Partitioning Around Medoids (PAM) clustering algorithm. The results were assessed for the optimal number of clusters using the Calinski-Harabasz (CH) Index. Between-class analysis (BCA) was performed to support the clustering and identify the drivers for the enterotypes. The analysis was done with the ade4 package. Prior to this analysis, in the dataset, OTUs with very low abundance were removed to decrease the noise, if their average abundance across all samples was below 0.01%.

2.4 Differential abundance analysis

DESeq2 package was used to normalize the OTUs table with the possible confounded factors and to detect OTUs that show significantly differential relative abundance between 2 assessed groups per comparison (with mental illness and without). OTUs with adjusted p-values < 0.05 and absolute estimated fold change > 4 were considered significantly differentially abundant between 2 examined classes.

2.5 Machine learning algorithms

Data was split in train and test dataset containing 80 and 20 percent each of the total data. createDataPartition function from the caret package was used for this purpose. Seven machine learning algorithms were trained using caret package; Random Forest (RF), Linear discriminant analysis (LDA), Support vector machine (SVM), K-nearest neighbor (KNN), Artificial neural network (ANN), Gradient boosting (GB) and Extreme gradient boosting (XGB). From the package foreach, a parallelized function was used in each process. The train control was made it with repeated cross validation with 10 partitions and it was repeated 5 times. The hyperparameters for tuning each algorithm can be seen in the GitHub repository (https://github.com/EduardoHerreros/Master-Thesis). The methods used in each algorithm are the followers: RF: ranger, LDA: lda, KNN: knn, SVM: svmRadial, ANN: nnet, GB: gbm and XGB: xgbTree. Accuracy was the metric that defined the optimal model in each algorithm. Models were compared between them using the average accuracy. Model's predictive potential were tested using the test dataset and compared between each other with the accuracy metric.

3. Results

3.1 Baseline characteristics of the study population.

In this project, the data of 16569 persons was included. From those, 7276 had complete cases. After we applied the eligibility criteria explained in methods, 5910 left. Finally, we chose randomly the same number of samples with and without mental illness, obtaining a final number of 2322 of gut microbiome samples with the metadata information. All continues variables presented a non-normal distribution (p-values < 0.05), calculated by the Kolmogorov-Smirnov normality test. The population characteristics are summarized in table1, categorical variables were analyzed using chi-square and continuous variables using Mann Whitney U test. Statistical differences regarding the mental illness states were found. As they can act as possible confounders factors, DESeq2 analysis and machine learning algorithms were corrected by these factors, excepts weight, because no differences were found in the BMI.

Table 1: Study population characteristics regarding the mental illness state

	Mental Illness		
Characteristic	Yes (n=1161)	No (n=1161)	p.value
Age (y)	44.0 (34.0;57.0)	49.0 (36.0;61.0)	**<0.001**
Alcohol consumption	930 (80.1)	1009 (86.9)	**<0.001**
Antibiotic	258 (22.2)	186 (16.2)	**<0.001**
Appendix removed	135 (11.6)	127 (10.9)	0.65
BMI	23.1 (21.2;25.5)	23.6 (21.5;25.5)	0.11
Country	1161 (50.0)	1161 (50.0)	0.08
Diabetes	21 (1.8)	14 (1.2)	0.31
Diet type			**<0.001**
- Omnivore	881 (75.9)	945 (81.4)	
- Omnivore not red meat	74 (6.4)	66 (5.7)	
- Vegan	64 (5.5)	27 (2.3)	
- Vegetarian	53 (4.6)	58 (5.0)	
- Vegetarian yes seafood	89 (7.7)	65 (5.6)	
Height (cm)	172.0 (164.0;177.0)	172.0 (165.0;179.0)	0.12
Inflammatory bowel disease	97 (8.4)	102 (8.8)	0.77
Irritable bowel syndrome	368 (31.7)	216 (18.6)	**<0.001**
Migraine	297 (25.6)	154 (13.3)	**<0.001**
Probiotic	744 (64.1)	621 (53.5)	**<0.001**
Ethnicity			**0.04**
- African American	5 (0.4)	7 (0.6)	
- Asian	32 (2.8)	58 (5.0)	
- Caucasian	1072 (92.3)	1036 (89.2)	
- Hispanic	20 (1.7)	23 (2.0)	
- Other	32 (2.8)	37 (3.2)	
Sex (male)	496 (42.7)	564 (48.8)	**0.004**
Small Intestinal Bacterial Overgrowth	113 (9.7)	51 (4.4)	**<0.001**
Weight (Kg)	68.0 (59.0;77.0)	70.0 (60.0;80.0)	**0.02**

3.2 Gut microbiota composition in the study population.

Differences in the relative abundance of bacteria have been observed at phylum (Fig. 2), family (Fig. 3) and genus (Fig. 4) level. At phylum level, *Bacteroidetes*, *Firmicutes*, *Proteobacteria* are found to be significantly different ($p < 0.05$) among the mental illness states. In the genus level, *Bacteroides* is more abundant in those people with one mental illness ($p = 0.007$). The Firmicutes/Bacteroidetes ratio was greater in people without mental illness than people with mental illness. *Firmicutes* abundance is higher and *Bacteroidetes* abundance is lower in people without mental illness issues.

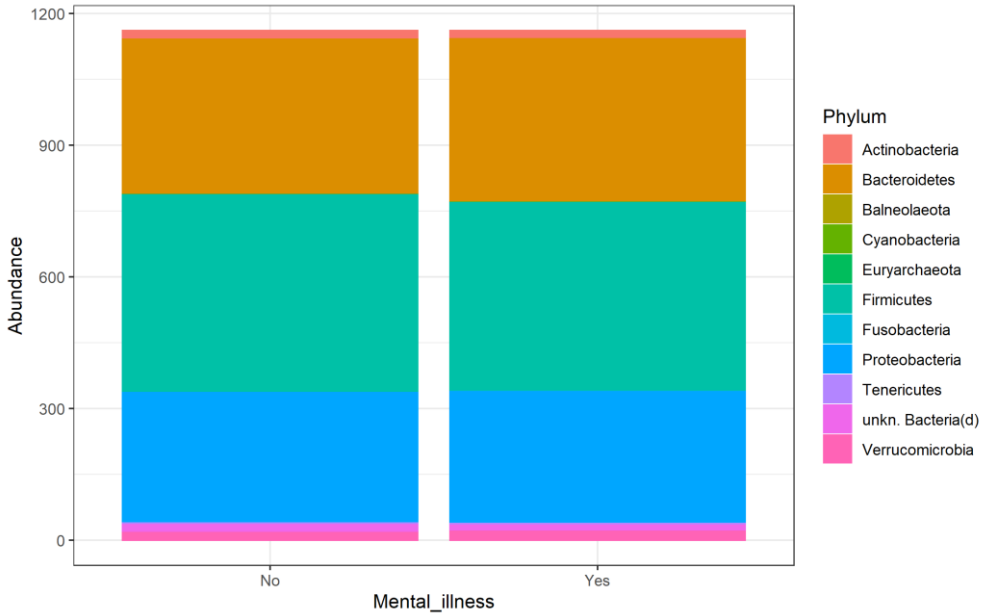

Figure 1: Phylum composition of bacteria from gut samples of people with or without mental illness.

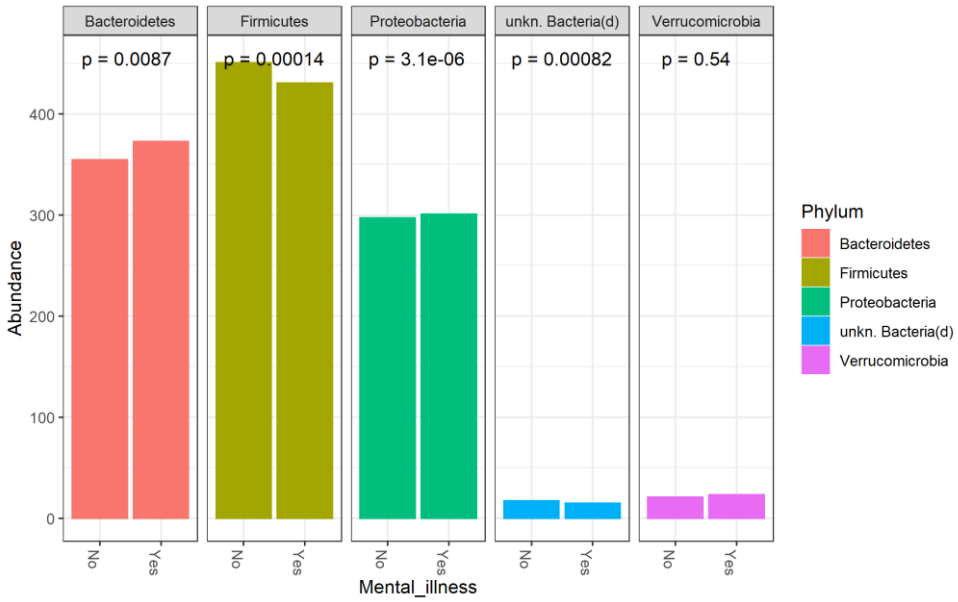

Figure 2: Top 5 Phylum order of gut samples of people with or without mental illness.

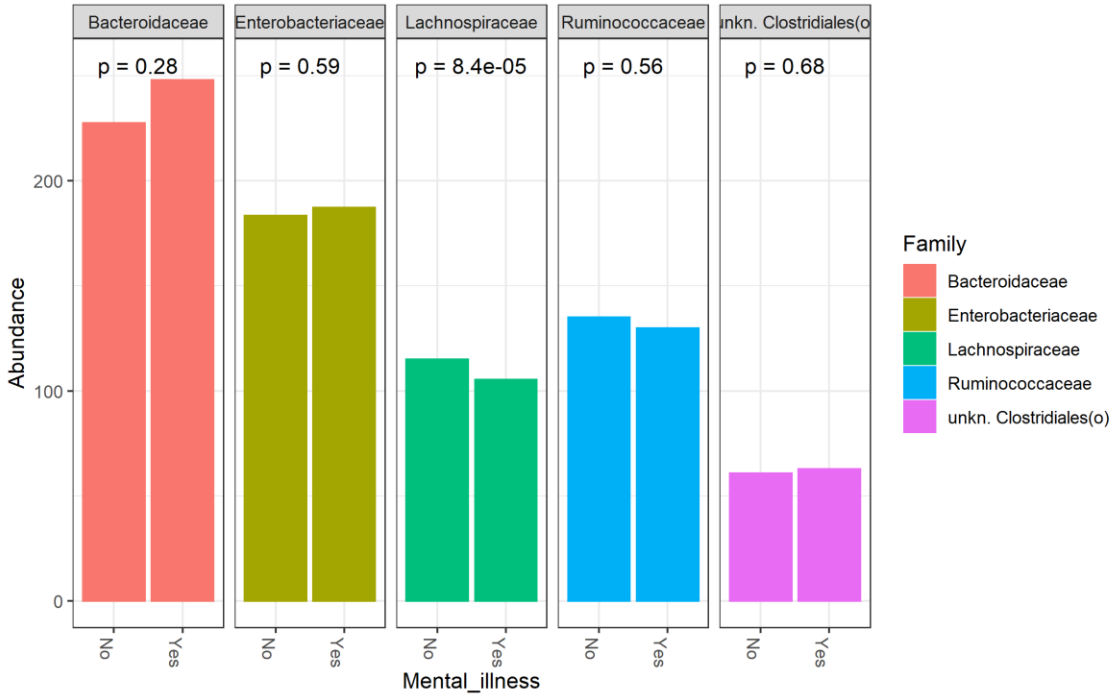

Figure 3: Top 5 Family order of gut samples of people with or without mental illness.

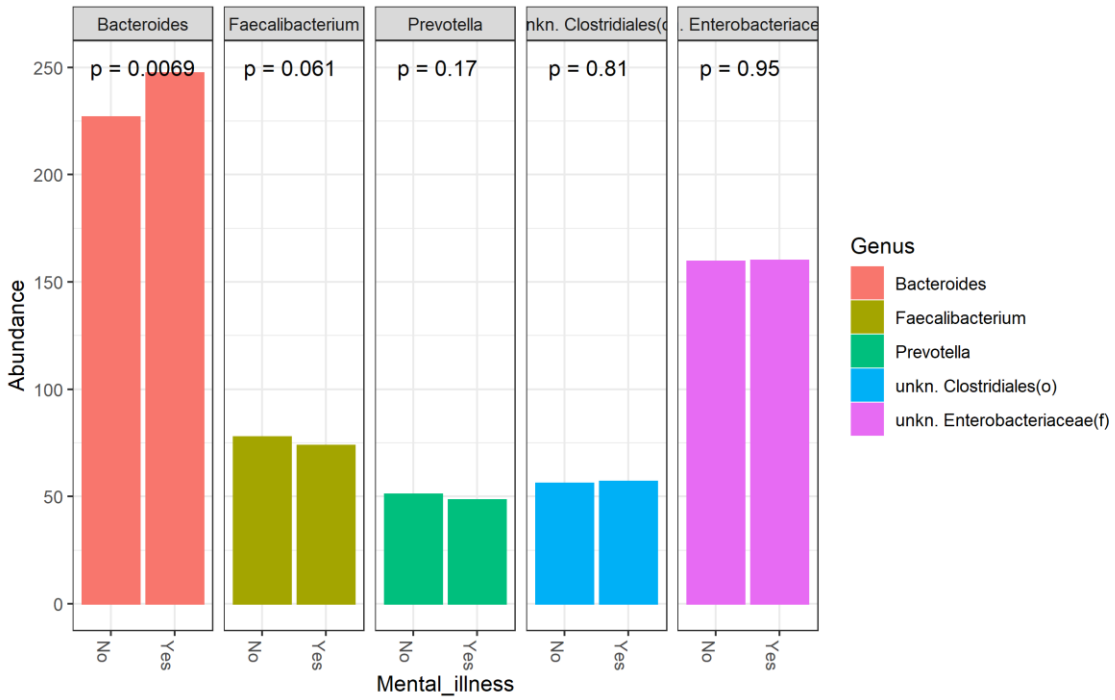

Figure 4: Top 5 Genus order of gut samples of people with or without mental illness.

3.3 Alpha and Beta-diversity of the gut microbiota in mental illness population.

People with some mental illness problem had lower microbial diversity (p = 0.002, Shannon index) but no differences in microbial richness (Fig. 5) were found. Beta diversity analysis using Bray Curtis distance do not showed visually differences between clusters regarding the mental illness states (Fig. 6). However, adonis test (PERMANOVA) was significant (p = 0.032) so the null hypothesis that mental illness states have the same centroid can be rejected. Moreover, the betadisper results are not significant (p = 0.849), it means that the null hypothesis that our groups have the same dispersions can be reject. This strengthen the adonis results, that is not due to differences in between groups dispersion.

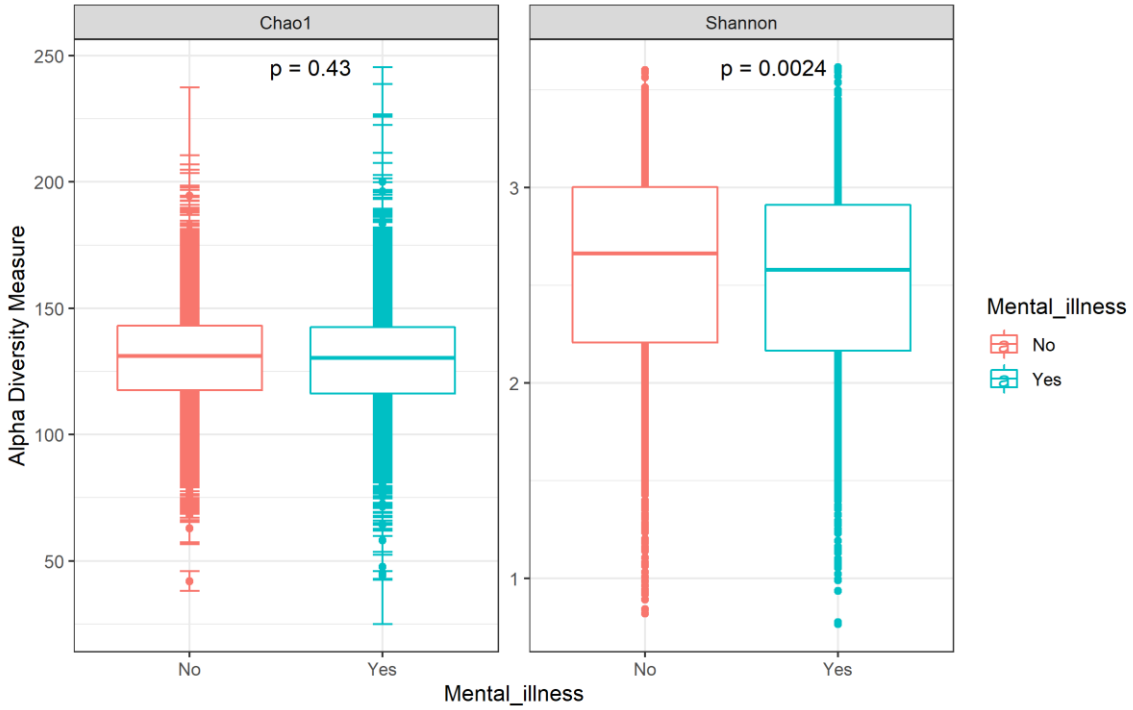

Figure 5: Boxplots of diversity indices for the gut microbiome, Chao1 and Shannon indices.

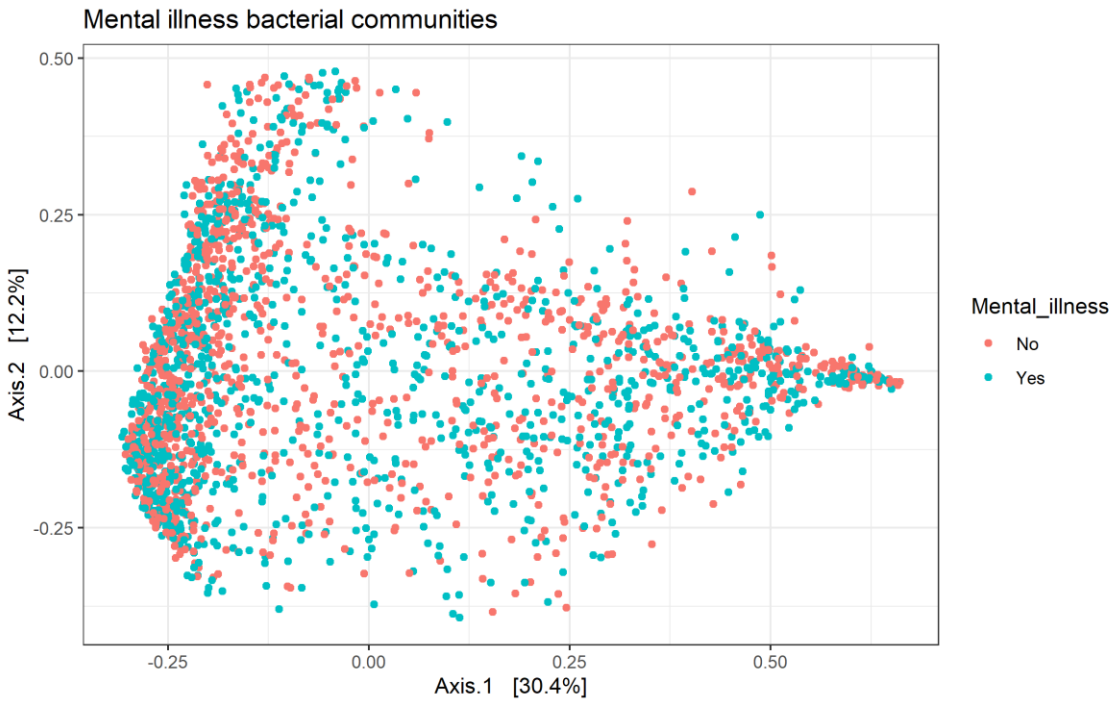

Figure 6: Ordination (PCoA) generated by using the Bray–Curtis dissimilarity metric for the two mental illness states. Samples are colored according to mental illness state.

3.4 Enterotype of the gut microbiota of the study population

In the enterotype analysis, 2 cluster were found as the optimal number of clusters in the study population (Fig. 7). The driver taxa for the cluster's formation were *Bacteroidaceae* and *Enterobacteriaceae*, OTU410 and OTU2503 respectively (Fig. 8).

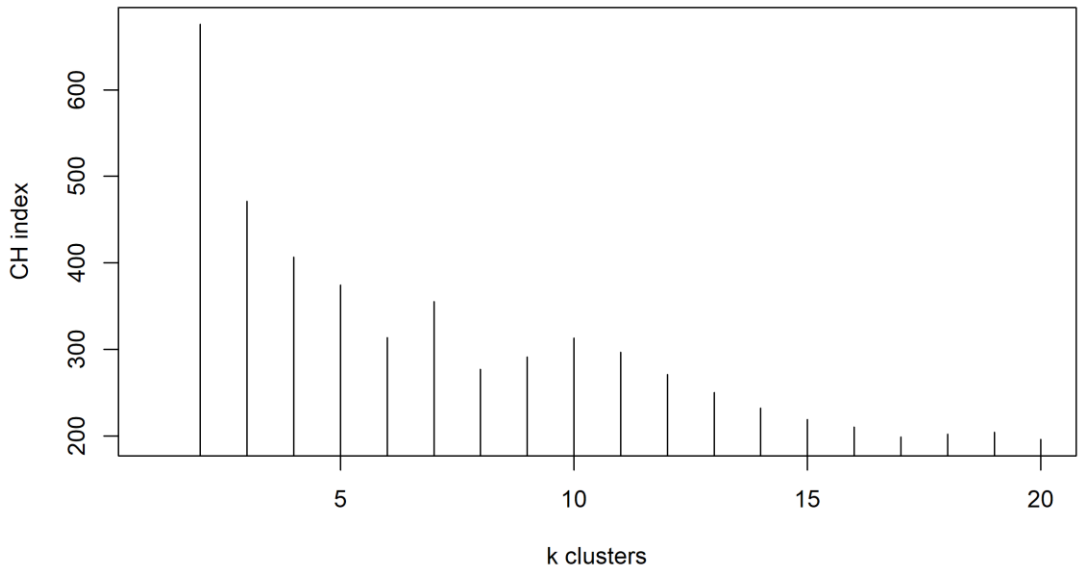

Figure 7: Optimal clusters number.

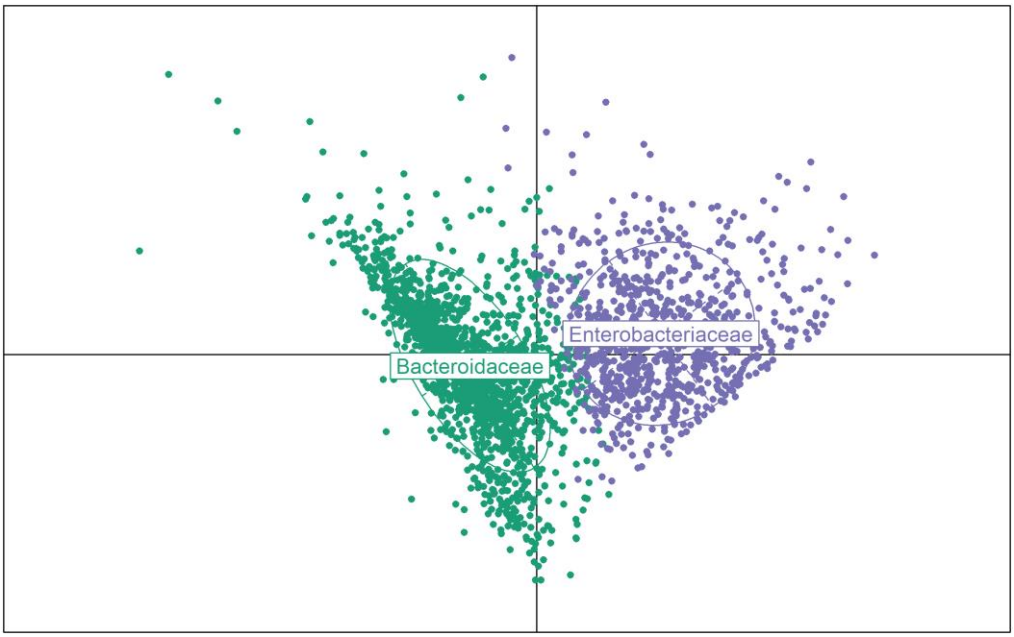

Figure 8: Between-class analysis (BCA) of the enterotypes.

3.5 Differential abundance analysis of the gut microbiota

To explore the differential composition of the gut microbiota in response to the different mental illness states DESeq2 analysis were performed. Fifty-one OTUs were significantly different (p-adj value < 0.05) regarding the mental states (Table 2). Only 4 of these 51 also had the estimate fold change superior to 4. These OTUs were: OTU2761 (*Proteobacteria*) higher abundance in healthy people, OTU1513 (*Firmicutes*) higher abundance in people with mental illness, OTU440 (*Bacteroidetes*) higher abundance in people with mental illness and OTU2325 (*Proteobacteria*) higher abundance in healthy people. The phylum *Bacteroidetes* (blue dots) is differentially expressed in only those who had one mental illness (Fig. 9). On the other side, *Proteobacteria* (red dots) tends to be more expressed in those people without mental illness.

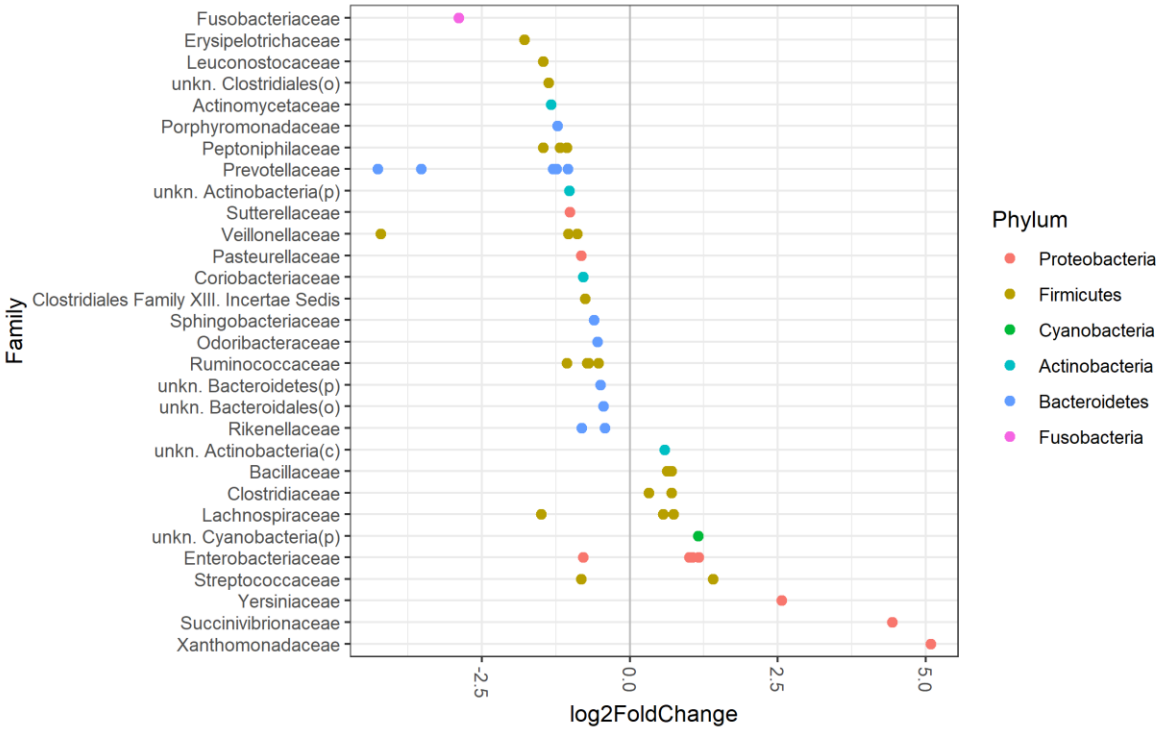

Figure 9: Phylum and family order representation of the DESeq2 results.

Table 2: Significantly enriched OTUs in the gut microbiome regarding mental illness state

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	Domain	Phylum	Class	Order	Family	Genus
OTU2545	13.705943	2.5663056	0.3630370	7.068992	0.0000000	0.0000000	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacterales	Yersiniaceae	unkn. Yersiniaceae(f)
OTU2761	87.106518	5.0814893	0.7204595	7.053122	0.0000000	0.0000000	Bacteria	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	Stenotrophomonas
OTU1513	14.809982	-4.2103375	0.6354025	-6.626253	0.0000000	0.0000000	Bacteria	Firmicutes	Negativicutes	Veillonellales	Veillonellaceae	Megasphaera
OTU974	3.021751	1.1556007	0.1815472	6.365292	0.0000000	0.0000000	Bacteria	Cyanobacteria	unkn. Cyanobacteria(p)	unkn. Cyanobacteria(p)	unkn. Cyanobacteria(p)	unkn. Cyanobacteria(p)
OTU1547	14.335077	-2.8925528	0.4999357	-5.785849	0.0000000	0.0000003	Bacteria	Fusobacteria	Fusobacteriia	Fusobacteriales	Fusobacteriaceae	Fusobacterium
OTU436	6.390752	-1.2206104	0.2144104	-5.692869	0.0000000	0.0000004	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Porphyromonadaceae	unkn. Porphyromonadaceae(f)
OTU1350	3.648208	-1.0680222	0.2074954	-5.147209	0.0000003	0.0000069	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Anaeromassilibacillus
OTU440	3.929710	-4.2576196	0.8334243	-5.108586	0.0000003	0.0000073	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Metaprevotella
OTU444	65.610890	-1.2954405	0.2544984	-5.090171	0.0000004	0.0000073	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	unkn. Prevotellaceae(f)
OTU1473	5.346837	-1.7823033	0.3623195	-4.919147	0.0000009	0.0000160	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Turicibacter
OTU106	3.360678	-1.3282012	0.2742263	-4.843450	0.0000013	0.0000213	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Actinomyces
OTU2325	15.625081	4.4350686	0.9403341	4.716482	0.0000024	0.0000368	Bacteria	Proteobacteria	Gammaproteobacteria	Aeromonadales	Succinivibrionaceae	Succinivibrio
OTU394	3.626491	-1.0255931	0.2214776	-4.630686	0.0000036	0.0000516	Bacteria	Actinobacteria	unkn. Actinobacteria(p)	unkn. Actinobacteria(p)	unkn. Actinobacteria(p)	unkn. Actinobacteria(p)
OTU443	30.384267	-3.5272358	0.7985175	-4.417231	0.0000100	0.0001226	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotellamassilia
OTU1523	20.460148	-1.4659694	0.3311534	-4.426859	0.0000096	0.0001226	Bacteria	Firmicutes	Tissierellia	Tissierellales	Peptoniphilaceae	Finegoldia
OTU1242	2.893069	-0.7569362	0.1802399	-4.199604	0.0000267	0.0003075	Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiales Family XIII. Incertae Sedis	unkn. Clostridiales Family XIII. Incertae Sedis(f)
OTU1220	8.921912	0.7073484	0.1740525	4.063995	0.0000482	0.0004931	Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	Lactonifactor
OTU1376	414.671896	-0.5279564	0.1298702	-4.065263	0.0000480	0.0004931	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcus
OTU1194	52.582308	-0.8220904	0.2086319	-3.940386	0.0000814	0.0007878	Bacteria	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus
OTU442	1233.553057	-1.0505317	0.2681122	-3.918255	0.0000892	0.0008206	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella
OTU1396	31.978763	-1.3725276	0.3678714	-3.730998	0.0001907	0.0016711	Bacteria	Firmicutes	Clostridia	Clostridiales	unkn. Clostridiales(o)	Fenollaria
OTU1362	130.359399	-0.6937299	0.1904850	-3.641914	0.0002706	0.0022634	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Gemmiger
OTU1030	59.400734	0.7049505	0.2016598	3.495741	0.0004727	0.0037820	Bacteria	Firmicutes	Bacilli	Bacillales	Bacillaceae	Bacillus
OTU344	13.514577	-0.7903388	0.2270414	-3.481034	0.0004995	0.0038294	Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Coriobacteriaceae	Collinsella
OTU1296	3.079355	-1.4965583	0.4391963	-3.407493	0.0006556	0.0048254	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Moryella
OTU1516	24.730575	-0.8889253	0.2631873	-3.377539	0.0007314	0.0051759	Bacteria	Firmicutes	Negativicutes	Veillonellales	Veillonellaceae	Veillonella
OTU1377	37.324735	-0.7210192	0.2155540	-3.344959	0.0008229	0.0056082	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruthenibacterium
OTU2484	68.065785	1.1605711	0.3572600	3.248533	0.0011600	0.0076230	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacterales	Enterobacteriaceae	Enterobacter
OTU1529	16.078092	-1.1832730	0.3672680	-3.221824	0.0012738	0.0080819	Bacteria	Firmicutes	Tissierellia	Tissierellales	Peptoniphilaceae	Peptoniphilus
OTU732	140.204872	-0.4968297	0.1561885	-3.180961	0.0014679	0.0090030	Bacteria	Bacteroidetes	unkn. Bacteroidetes(p)	unkn. Bacteroidetes(p)	unkn. Bacteroidetes(p)	unkn. Bacteroidetes(p)
OTU1192	5.698367	-1.4665257	0.4749151	-3.087975	0.0020153	0.0119615	Bacteria	Firmicutes	Bacilli	Lactobacillales	Leuconostocaceae	Weissella
OTU339	13.684110	0.5897599	0.1944206	3.033423	0.0024180	0.0139033	Bacteria	Actinobacteria	Actinobacteria	unkn. Actinobacteria(c)	unkn. Actinobacteria(c)	unkn. Actinobacteria(c)
OTU2041	44.888111	-1.0140852	0.3389626	-2.991732	0.0027740	0.0154671	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Sutterellaceae	Sutterella
OTU419	25.860479	-0.5468352	0.1844700	-2.964358	0.0030332	0.0160905	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Odoribacteraceae	Odoribacter
OTU2496	17.302400	1.0575610	0.3570940	2.961576	0.0030607	0.0160905	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacterales	Enterobacteriaceae	Raoultella
OTU1193	3.950237	1.4026194	0.4834377	2.901345	0.0037156	0.0188763	Bacteria	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Lactococcus
OTU2649	23.010358	-0.8224278	0.2841197	-2.894653	0.0037958	0.0188763	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus
OTU441	110.815742	-1.2401327	0.4341805	-2.856261	0.0042866	0.0207563	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Paraprevotella
OTU2489	25.263839	1.0032250	0.3527079	2.844351	0.0044502	0.0209958	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacterales	Enterobacteriaceae	Klebsiella
OTU1521	12.001016	-1.0671981	0.3772653	-2.828774	0.0046727	0.0214943	Bacteria	Firmicutes	Tissierellia	Tissierellales	Peptoniphilaceae	Anaerococcus
OTU1515	2.785564	-1.0382081	0.3694974	-2.809785	0.0049575	0.0222481	Bacteria	Firmicutes	Negativicutes	Veillonellales	Veillonellaceae	unkn. Veillonellaceae(f)
OTU453	16.285029	-0.8172272	0.2933813	-2.785547	0.0053438	0.0234107	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae	unkn. Rikenellaceae(f)
OTU1231	65.122446	0.3241259	0.1168631	2.773552	0.0055448	0.0237266	Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	unkn. Clostridiaceae(f)
OTU1078	11.476490	0.6267621	0.2287405	2.740058	0.0061428	0.0256882	Bacteria	Firmicutes	Bacilli	Bacillales	Bacillaceae	unkn. Bacillaceae(f)
OTU1283	4.452088	0.5626551	0.2060706	2.730400	0.0063258	0.0258653	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Faecalicatena
OTU458	177.681765	-0.4445377	0.1645708	-2.701195	0.0069091	0.0276363	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	unkn. Bacteroidales(o)	unkn. Bacteroidales(o)
OTU1370	84.818464	-0.5281020	0.1981596	-2.665033	0.0076981	0.0301371	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Oscillospira
OTU2483	12.770190	-0.7910395	0.3016857	-2.622065	0.0087399	0.0335028	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacterales	Enterobacteriaceae	Cronobacter
OTU446	347.043826	-0.4214386	0.1636261	-2.575620	0.0100061	0.0375738	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae	Alistipes
OTU719	6.218431	-0.6090388	0.2423970	-2.512568	0.0119856	0.0434063	Bacteria	Bacteroidetes	Sphingobacteriia	Sphingobacteriales	Sphingobacteriaceae	Sphingobacterium
OTU1286	8.554717	0.7384414	0.2940556	2.511231	0.0120311	0.0434063	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Hespellia

3.6 Mental illness prediction by machine learning algorithms.

Seven machine learning methods were performed to achieved one model capable to predict the predisposition to some mental illness regarding the gut microbiota. Multiple models were created with different hyperparameters trying to look for the best model in each ML method. A 10-fold cross validation was applied to each method. No model come out as a good predictor using the gut microbiota information, the worst model presented a mean accuracy of 0.5 (ANN) and the best one, Random Forest, of 0.61 (Fig. 10). When the test data where used for probing the model prediction, Random forest was the model with the better performance (Accuracy = 0.62), and the worst KNN with 0.48 (Fig. 11).

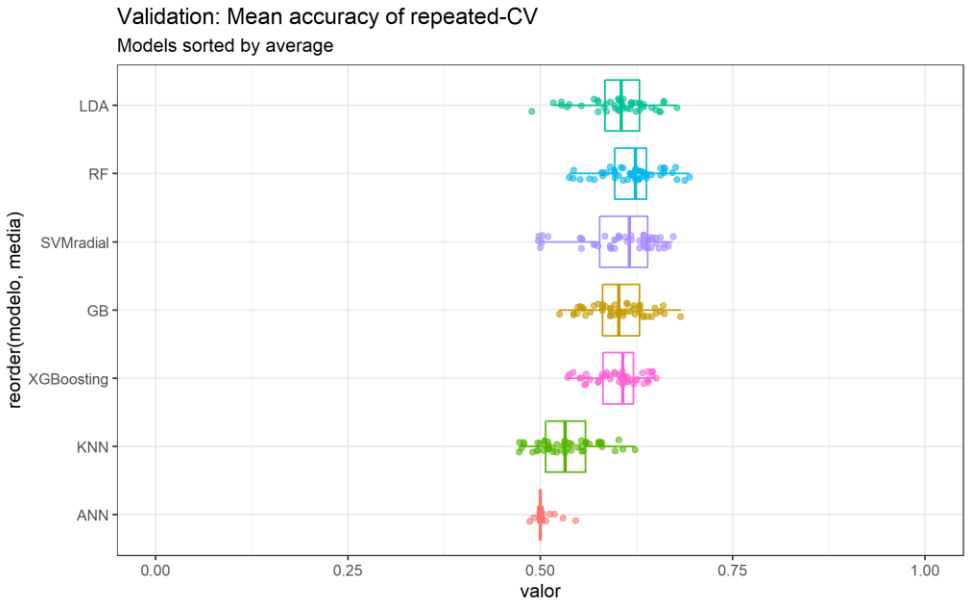

Figure 10: Predictive accuracy comparison of the machine learning models.

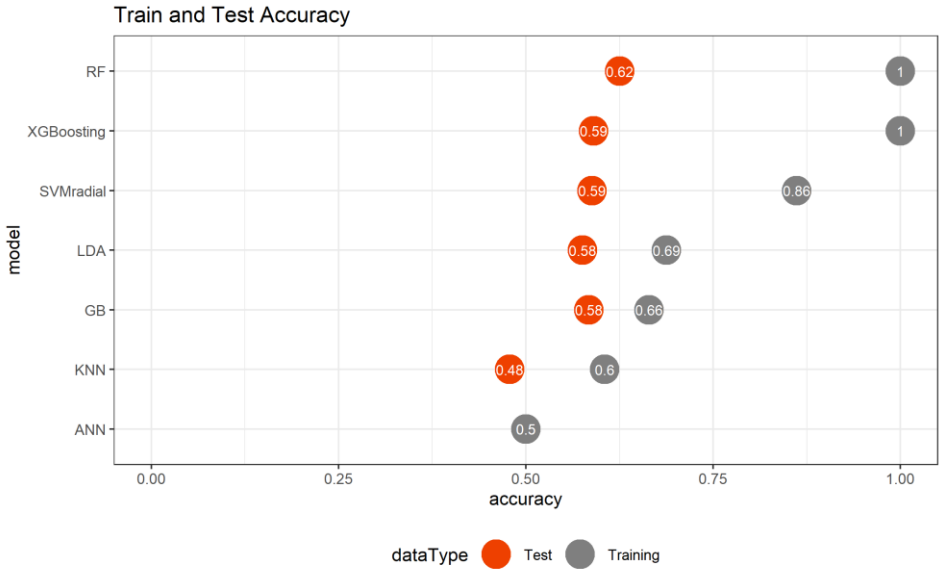

Figure 11: Machine learning models prediction accuracy using the training and test dataset

4. Discussion

In this study, we have shown that the gut microbiota of individuals with one mental illness is distinct from those people without a mental issue and it is reflected in the gut microbiota abundance profiles. In the demographic analyses we found interesting results regarding some features typically linked with changes in the microbiome, but not necessarily associated with mental health issues. The irritable bowel syndrome (IBS), and small intestinal bacterial overgrowth (SIBO) are pathologies associated with the unbalance of gut microbiota composition. In people with mental illness issues we found an incidence ratio of 2:1 in the SIBO and IBS when it is compared to people without mental issues. SIBO is an excess of bacteria, sometimes methane producing bacteria and more often with hydrogen producing bacteria in the small intestine in a portion of the digestive system that normally has few bacterial or other microbial residents. IBS is a common functional bowel disorder characterized by abdominal pain, cramping, bloating, gas, diarrhea, constipation. The research community links IBS and SIBO because of the similarity of the symptoms with some believing that IBS is the primary event predisposing to SIBO, when others believe just the opposite (17, 18). We found that the ethnicity rates are different in those people with mental illness than in the healthy ones. In our results, Asians have a 50% lower incidence of mental problems, also women are more associated to suffer one mental health problem. This is in line with the National Institute of mental health (US) (19). We found that the diet type can be a risk factor to a mental illness issue. Vegan diet shows to be the main diet type that presented differences among groups. Lavalle et al found a slightly increases over time in anxiety and depression in Chinese student with vegan diet, however in Germans, Russians or Americans students this wasn't found. To our knowledge, there are not studies linking plant-based diets and cognitive abilities on a neural level, which are urgently needed, due to the hidden potential as a dietary therapeutic tool (20, 21).

In our study, healthy people had higher microbial diversity and distinct complexity, showed by the ordination and PERMANOVA analysis. In line with these results, clinical evidence in psychiatric disorders also associate an alteration in microbiota diversity and complexity when compared to healthy controls as determined for autism (22, 23), attention deficit (5), schizophrenia and bipolar disorders (24), Alzheimer (25) and other mood and neuroinflammatory linked disorders (26-28).

Bacteroidetes in our work is in greater abundance in people with mental illness that in healthy people, and *Firmicutes* in higher abundance in healthy people. This is supported by Liu et al who found an enrichment of *Bacteroidetes* and reduction of *Firmicutes* in people with depression. They also highlight the importance of the IBS in this pathology, our work found that mental illness people also have higher indices of IBS in their group (27). The Firmicutes/Bacteroidetes ratio was higher in people without mental illness. The same patron has been observed in previously works describing an association with a decrease in the levels of *Firmicutes* and increase in *Bacteroidetes* levels (18, 29).

Regarding the enterotype analysis, published by Arumugam, Raes et al, we identified two groups and not 3 drivers genera as they found. Ours driver's genera were *Bacteroides* and an unknown *Enterobacteriaceae*. In our microbiota composition results, *Bacteroides* are in higher abundance in people with mental illness, but the *Enterobacteriaceae* family did not show differences between groups. Arumugam, Raes et al and Liu et al also found that Bacteroidetes was one of the driver genera for the enterotypes classification. In contrast with their results, we did not find that *Prevotella* to be one of the other driver genera. The differential analysis performed using DESeq2 showed that Bacteroidetes are in higher proportion in the phylum level in those people with mental issues. This is consistent with the gut microbial composition and the enterotype analysis. There are publications mentioning the *Enterobacteriaceae* family as an enterotype in some populations. In addition, a recent article suggested enterotypes might not even exist, as by increasing the number of samples, no clear clusters are seen (30).

To evaluate the prediction power of the gut microbiota to some mental issue, we trained 7 machine learning methods. The best predictor model was RF with an accuracy of 62% of the test dataset. However, as our data is weighted with the same number of people with and without mental illness, the prediction power of our models is not greater than randomly predicted its state without knowing any other characteristic (50% right, 50% wrong). In the last years, a whole new window has been open regarding the microbiome studies and the power of machine learning algorithms (15, 31-34). Existing studies often report disease-associated dysbiosis, a microbial imbalance inside the host, but such associations can have a wide range of interpretations. To our knowledge, there are not studies of machine learning models using the gut microbiome information with any mental health issue. There are several limitations in the microbiome studies and the use of machine learning algorithms. First, OTU tables are sparse, with a large proportion of zero counts (35). Second, the role of taxonomy in prediction is often unclear – similar sequences are often correlated across samples, which is a property that can be readily assessed directly without taxonomic knowledge. Third, library sizes (essentially column sums of the OTU table) vary considerably, and normalization methods must be used to account for this variation (36). In our case, the mixed of different mental disorders with a different biological background could be making noise and be the explanation why our ML models do not have one better prediction results. More efforts are necessary in the use of machine learning algorithms with microbiome information, because of the potential that these methods have in the classification and/or prediction of certain pathologies.

5. Conclusions

- A lower gut diversity (Shannon index) was observed in people with one mental illness issue.

- Mental illness states had different centroid in the ordination plot of the beta diversity analysis.

- A lower *Bacteroidetes* and higher *Firmicutes* abundance were found in people without mental problems.

- *Bacteroides* and an unknown *Enterobacteriaceae* were the driver genera of the microbiota of the study.

- None of the machine learning models presented a good prediction accuracy for the mental illness states regarding the gut microbiota information.

6. Glossary

Microbiome: the genome of all our microbes.

Microbiota: wide collection of microbes which conforms a microbial community.

Alpha diversity: the average species diversity in a habitat or specific area.

Beta diversity: the ratio between local or alpha diversity and regional diversity. This is the diversity of species between two habitats or regions

Enterotypes: way to stratify human individuals based on their gut microbiome.

Principal Component Analysis (PCoA): method to explore and to visualize similarities or dissimilarities of data.

K-Nearest Neighbor: It is based on the idea of identifying observations in the training set that resemble the observation of tests (neighboring observations) and assign the predominant class among those observations as the predicted value.

Linear Discriminant Analysis: Supervised classification method of qualitative variables in which two or more groups are known a priori and new observations are classified into one of them according to their characteristics.

Random Forest: It is a modification of the bagging process that achieves better results because it decorates the trees generated in the process. Bagging is based on the fact that, by averaging a set of models, the variance is reduced.

Support Vector Machine: They are based on the Maximal Margin Classifier, which in turn is based on the concept of hyperplane.

Artificial Neural Network: is an information processing paradigm that is inspired by the way the biological nervous system such as brain process information. It is composed of large number of highly interconnected processing elements (neurons) working in unison to solve a specific problem

Gradient Boosting: is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

Extreme Gradient Boosting: is one of the implementations of Gradient Boosting concept, but XGB uses a more regularized model formalization to control over-fitting, which gives it better performance.

7. References

1. NHS England. Mental Health Taskforce for NHS England The Five Year Forward View for Mental Health: A Report from the Independent Mental Health Taskforce to the NHS in England. 2017 [Available from: https://www.england.nhs.uk/wp-content/uploads/2017/03/fyfv-mh-one-year-on.pdf.
2. Kastrup M. Global burden of mental health: Springer; 2010.
3. World Health Organization. Mental Health Included in the UN Sustainable Development Goals [Available from: https://www.who.int/mental_health/SDGs/en/.
4. Organization WH. Third United Nations High-level Meeting on NCDs [
5. Chandra PS, Chand P. Towards a new era for mental health. Lancet. 2018;392(10157):1495-7.
6. Patel V, Saxena S, Lund C, Thornicroft G, Baingana F, Bolton P, et al. The Lancet Commission on global mental health and sustainable development. Lancet. 2018;392(10157):1553-98.
7. Bravo JA, Julio-Pieper M, Forsythe P, Kunze W, Dinan TG, Bienenstock J, et al. Communication between gastrointestinal bacteria and the nervous system. Curr Opin Pharmacol. 2012;12(6):667-72.
8. Clapp M, Aurora N, Herrera L, Bhatia M, Wilen E, Wakefield S. Gut microbiota's effect on mental health: The gut-brain axis. Clin Pract. 2017;7(4):987.
9. Lynch SV, Pedersen O. The Human Intestinal Microbiome in Health and Disease. N Engl J Med. 2016;375(24):2369-79.
10. Bloomfield SF, Rook GA, Scott EA, Shanahan F, Stanwell-Smith R, Turner P. Time to abandon the hygiene hypothesis: new perspectives on allergic disease, the human microbiome, infectious disease prevention and the role of targeted hygiene. Perspect Public Health. 2016;136(4):213-24.
11. Konig J, Wells J, Cani PD, Garcia-Rodenas CL, MacDonald T, Mercenier A, et al. Human Intestinal Barrier Function in Health and Disease. Clin Transl Gastroenterol. 2016;7(10):e196.
12. Fond G, Boukouaci W, Chevalier G, Regnault A, Eberl G, Hamdani N, et al. The "psychomicrobiotic": Targeting microbiota in major psychiatric disorders: A systematic review. Pathol Biol (Paris). 2015;63(1):35-42.
13. Spadoni I, Zagato E, Bertocchi A, Paolinelli R, Hot E, Di Sabatino A, et al. A gut-vascular barrier controls the systemic dissemination of bacteria. Science. 2015;350(6262):830-4.
14. Knights D, Costello EK, Knight R. Supervised classification of human microbiota. FEMS Microbiol Rev. 2011;35(2):343-59.
15. Moitinho-Silva L, Steinert G, Nielsen S, Hardoim CCP, Wu YC, McCormack GP, et al. Predicting the HMA-LMA Status in Marine Sponges by Machine Learning. Front Microbiol. 2017;8:752.
16. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, et al. Enterotypes of the human gut microbiome. Nature. 2011;473(7346):174-80.
17. Rao SSC, Rehman A, Yu S, Andino NM. Brain fogginess, gas and bloating: a link between SIBO, probiotics and metabolic acidosis. Clin Transl Gastroenterol. 2018;9(6):162.

18. Labus JS, Hollister EB, Jacobs J, Kirbach K, Oezguen N, Gupta A, et al. Differences in gut microbial composition correlate with regional brain volumes in irritable bowel syndrome. Microbiome. 2017;5(1):49.
19. National Institute of Mental Health. Any Mental Illness Among US Adults [Available from: https://www.nimh.nih.gov/health/statistics/mental-illness.shtml.
20. Lavallee K, Zhang XC, Michalak J, Schneider S, Margraf J. Vegetarian diet and mental health: Cross-sectional and longitudinal analyses in culturally diverse samples. J Affect Disord. 2019;248:147-54.
21. Medawar E, Huhn S, Villringer A, Veronica Witte A. The effects of plant-based diets on the body and the brain: a systematic review. Transl Psychiatry. 2019;9(1):226.
22. Parracho HM, Bingham MO, Gibson GR, McCartney AL. Differences between the gut microflora of children with autistic spectrum disorders and that of healthy children. J Med Microbiol. 2005;54(Pt 10):987-91.
23. Tomova A, Husarova V, Lakatosova S, Bakos J, Vlkova B, Babinska K, et al. Gastrointestinal microbiota in children with autism in Slovakia. Physiol Behav. 2015;138:179-87.
24. Dickerson F, Severance E, Yolken R. The microbiome, immunity, and schizophrenia and bipolar disorder. Brain Behav Immun. 2017;62:46-52.
25. Vogt NM, Kerby RL, Dill-McFarland KA, Harding SJ, Merluzzi AP, Johnson SC, et al. Gut microbiome alterations in Alzheimer's disease. Sci Rep. 2017;7(1):13537.
26. Hu S, Li A, Huang T, Lai J, Li J, Sublette ME, et al. Gut Microbiota Changes in Patients with Bipolar Depression. Adv Sci (Weinh). 2019;6(14):1900752.
27. Liu Y, Zhang L, Wang X, Wang Z, Zhang J, Jiang R, et al. Similar Fecal Microbiota Signatures in Patients With Diarrhea-Predominant Irritable Bowel Syndrome and Patients With Depression. Clin Gastroenterol Hepatol. 2016;14(11):1602-11 e5.
28. Scheperjans F, Aho V, Pereira PA, Koskinen K, Paulin L, Pekkonen E, et al. Gut microbiota are related to Parkinson's disease and clinical phenotype. Mov Disord. 2015;30(3):350-8.
29. Huang Y, Shi X, Li Z, Shen Y, Shi X, Wang L, et al. Possible association of Firmicutes in the gut microbiota of patients with major depressive disorder. Neuropsychiatr Dis Treat. 2018;14:3329-37.
30. Knights D, Ward TL, McKinlay CE, Miller H, Gonzalez A, McDonald D, et al. Rethinking "enterotypes". Cell Host Microbe. 2014;16(4):433-7.
31. Ryan FJ. Application of machine learning techniques for creating urban microbial fingerprints. Biol Direct. 2019;14(1):13.
32. Roguet A, Eren AM, Newton RJ, McLellan SL. Fecal source identification using random forest. Microbiome. 2018;6(1):185.
33. Zeevi D, Korem T, Zmora N, Israeli D, Rothschild D, Weinberger A, et al. Personalized Nutrition by Prediction of Glycemic Responses. Cell. 2015;163(5):1079-94.
34. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, et al. Best practices for analysing microbiomes. Nat Rev Microbiol. 2018;16(7):410-22.
35. Hu T, Gallins P, Zhou YH. A Zero-inflated Beta-binomial Model for Microbiome Data Analysis. Stat (Int Stat Inst). 2018;7(1).

36. Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome. 2017;5(1):27.