REPUBLIC OF TURKEY

ACIBADEM MEHMET ALİ AYDINLAR UNIVERSITY

INSTITUTE OF HEALTH SCIENCES

INVESTIGATION of THE HUMAN GUT MICROBIOME ALTERATIONS in MULTIPLE SCLEROSIS

ZEHRA HAZAL SEZER

MASTER THESIS

DEPARTMENT of MEDICAL BIOTECHNOLOGY

Supervisor

Prof. Dr. Osman Uğur Sezerman

Co-Supervisor

Dr. Orhan Özcan

ISTANBUL-2020

DECLERATION

I declare that this thesis study entitled "Investigation of The Human Gut Microbiome Alterations in Multiple Sclerosis" submitted to Acıbadem MAA University, Istanbul, Turkey for the award of the degree of Master of Medical Biotechnology. This study has written by me and not submitted for any previous degree. From planning to writing, I had no unethical behavior at any stage.

20.07.2020

Zehra Hazal Sezer

ACKNOWLEDGEMENTS

First, I would like to thank my supervisors Prof. Dr. Uğur Sezerman and co-advisor Dr. Orhan Özcan for their help and guidance that led me to finish my thesis step by step. I want to give special express my sincere gratitude to my advisor Prof. Uğur Sezerman for the continuous support of my MSc study and research, for his patience, motivation, enthusiasm, and immense knowledge.

I would like to thank Prof. Dr. Uğur Sezerman head of jury, and jury members; Prof. Dr. Aksel Siva, Prof. Dr. Tanıl Kocagöz, Dr. Orhan Özcan and Dr. Öznur Taştan for evaluation of thesis project and found me qualified for MSc degree.

I would also like to thank Elif Kılıç, Eray Şahin, Okan Soykam, who carefully helped to correct my thesis and gave me encouraging oral defense suggestions. A special thanks to awesome SEZERMAN lab members. During my master's years, they made me happy and I had fun and unforgettable times with them.

Lastly, the greatest thanks to my family; I would like to thank my parents and grandma who always did everything they could to be a beneficial person for society; to my brother who supported me in every way believable. And special thanks to my friends who are Selin Keleş, Çağlanur Rüzgar, Alex Plesko and Nezih Üzümcüoğlu, I would not be able achieve this without their emotional support. I 'm lucky to have such a great family and friends, and I'm sure I'm going to feel my whole life this way.

TABLE OF CONTENTS

DECLERATION ..... **iii**
ACKNOWLEDGEMENTS ..... **iv**
TABLE OF CONTENTS ..... **v**
LIST OF ABBREVIATIONS and SYMBOLS ..... **vii**
LIST OF FIGURES ..... **viii**
SUMMARY **1**
ÖZET **2**
1.BACKGROUND AND AIM OF STUDY **3**
2. INTRODUCTION **5**
2.1 Multiple Sclerosis **5**
2.1.1 Genotypic Effects 6
2.1.2 Epigenetics Effects 7
2.1.3 Environmental Effects 8
2.2 Human Microbiome 9
2.2.1 Gut-Brain axis and Microbiome 10
2.2.2 Disease and Microbiome Alteration 11
2.3 Bacterial Diversity and Taxonomic Classification 14
2.3.1 Metabolic Diversity of Microorganisms 14
2.3.2 Bacterial Taxonomic Classifications 17
2.4 Machine Learning Algorithms 20
2.4.1 Decision Tree 21
2.4.2 Random Forest 22
3. MATERIALS AND METHODS **23**
3.1 Metagenomic Datasets 23
3.2 Taxonomic Analyses and Cluster Quality 23
3.3 Predictive Modeling using Random Forests 24
3.4 Functional characterization 25
4. RESULTS **27**
4.1 Microbial Profiling of RRMS Changes at The Species Level 27
4.2 Microbial Profiling of RRMS Changes at The Genus Level 32
4.3 Random Forest 37
4.5 Functional Analysis 42
5. DISCUSSION and CONCLUSION **44**
6. REFERENCES **51**

CURRICULUM VITAE..... 62

LIST OF ABBREVIATIONS and SYMBOLS

AD	Alzheimer's disease
BA	Bile Acids
BLAST	Basic Local Alignment Search Tool
CNS	Central Nervous System
DNA	Deoxyribonucleic acid
EBV	Epstein–Barr virus
GABA	π-aminobutyric Acid
GI	Gastrointestinal tract
GWAS	Genome-wide Association Analysis
HLA	Human Leukocyte Antigen
IL2RA	Interleukin-2 receptor α gene
IL7RA	Interleukin-7 receptor α gene
KEGG	Kyoto Encyclopedia of Genes and Genomes
MS	Multiple Sclerosis
NGS	Next-Generation Sequencing
NIH	National Institutes of Health
OTUs	Operating Taxonomic Units
PCR	Polymerase Chain Reaction
PPMS	Primary Progressive MS
RFs	Random Forests
RNA	Ribonucleic acid
rRNA	Ribosomal RNA
RRMS	Relapsing-remitting MS
SCFAs	Short-chain Fatty Acids
SRA	Sequence Read Archive
SPMS	Secondary Progressive MS
WHO	World Health Organization

LIST OF FIGURES

Figure 2.1. Types of Multiple Sclerosis 6
Figure 2.2. Overview risk factor for Multiple Sclerosis 8
Figure 2.3 Approximately 1.5 kb 16S rRNA E. coli gene that displays the nine variable regions 18
Figure 3. Overview of the method 26
Figure 4.1. Gut microbiota of MS patients differs from control groups. Decision Tree for first data in species-level at 71% accuracy with KRAKEN2 result. C: Control, MS: Multiple Sclerosis percentages in overall concentration. 28
Figure 4.2. Gut microbiota of MS patients differs from control groups. Decision Tree for first data in species-level at 64% accuracy with BLASTn result 29
Figure 4.3. Decision Tree for second data in species-level at 71% accuracy with KRAKEN2 result.. 30
Figure 4.4. Decision Tree for second data in species-level at 91.6% accuracy with BLAST-N result.. 31
Figure 4.5. Decision Tree for first data in genus-level at 65% accuracy with KRAKEN2 result.. 33
Figure 4.6. Decision Tree for first data in genus-level at 48% accuracy with BLASTn result.. 34
Figure 4.7. Decision Tree for second data in genus-level at 88% accuracy with KRAKEN2 result.. 35
Figure 4.8. Decision Tree for second data in genus-level at 64% accuracy with BLASTn result. 36
Figure 4.9. Random Forest for first data in species-level at 60% accuracy with KRAKEN2 result. 38
Figure 4.10. Random Forest for second data in genus-level at 92% accuracy with KRAKEN2 result. 39
Figure 4.11. Random Forest for second data in genus-level at 75% accuracy with BLASTn result. 40
Figure 4.12. Random Forest for second data in species-level at 92% accuracy with BLAST-N result. 41

Figure 4.13. Decision Tree for first data in genus-level at 85% accuracy with GhostKoala result.. 43

SUMMARY

Multiple Sclerosis (MS) is a type of autoimmune disease of the central nervous system (CNS) and characterized by the damage in myelinated axons in varying degrees. Depending on the severity of the immune-mediated attacks, the neural dysfunctions become permanent. Both genetic and environmental factors contribute to the development and progression of the disease, resulting in diverse clinical presentations. Within the last decade, thanks to the advancements in next-generation sequencing technology, microbiome studies gained great progression and revealed a significant relationship with MS. In this thesis, it was aimed to investigate the presence of any gut microbiome alterations in patients suffering from MS, and examination of possible consequences of this change. For that purpose, 16S sequencing data of two different studies have been used to identify distinguish a healthy individual from the MS patients using several machine learning algorithms. A significant amount of increase was detected in bacterial genera of *Lysinibacillus, Blautia, Akkermansia, Ruthenibacterium* and *Anaerostipes* in samples collected from MS patients, and they were assigned as potential markers in the classification of disease and normal phenotypes with at least 87% accuracy. Empirical pathway analysis by performing two machine learning algorithms, decision tree and random forest, predicted 'amytrophic lateral sclerosis', 'lipopolysaccharide biosynthesis', and 'glutathione metabolism' pathways, which were shown to be enhanced in MS patients, from KEGG database. The obtained accuracy in this classification was 89%. In conclusion, the bacterial gut microbiome compositions and their potential target pathways were determined, and potential markers in the classification of MS patients and healthy controls were proposed. Further studies to link microbiome composition and immune system of the host may have a great impact on the reveal of the disease development and progression, and help to propose new strategies against MS.

Keywords: Gut Microbiome, Multiple Sclerosis, Microbiome Analysis, Pathway Analysis, Machine Learning

ÖZET

Multiple Sklerozda İnsan Bağırsak Mikrobiyom Değişikliklerinin İncelenmesi

Multiple Skleroz (MS) merkezi sinir sisteminin (CNS) miyelinli aksonlarda değişen derecelerde hasar ile karakterize edilen bir tür otoimmün hastalığıdır. Nöral fonksiyon bozuklukları, bağışıklık aracılı atakların şiddetine bağlı olarak, kalıcı hale gelir. Hem genetik hem de çevresel faktörler, çeşitli klinik bulgularla sonuçlanan hastalığın gelişimine ve ilerlemesine neden olur. Son on yıl içinde, yeni nesil sekanslama teknolojisindeki gelişmeler sayesinde mikrobiyom çalışmalarında büyük ilerleme kaydedilmiş ve MS ile arasındaki anlamlı ilişki ortaya çıkarılmıştır. Bu tezde, MS hastalarında bağırsak mikrobiyom değişikliklerinin varlığının araştırılması ve bu değişikliğin olası sonuçlarının incelenmesi amaçlanmıştır. Bu doğrultuda, makine öğrenme algoritmaları kullanılarak sağlıklı bireyleri MS hastalarından ayırt etmek için iki farklı çalışmanın 16S sekans verileri kullanılmıştır. MS hastalarından toplanan örneklerde *Lysinibacillus, Blautia, Akkermansia, Ruthenibacterium* ve *Anaerostipes* bakteri cinslerinde önemli miktarda artış saptanmış ve hastalık ile normal fenotiplerin sınıflandırılmasında en az %87 doğrulukla bulunan bakteriler potansiyel belirteçler olarak belirlenmiştir. Karar ağacı ve rastgele orman makine öğrenme algoritmaları ile gerçekleştirilen ampirik yolak analizlerinde MS hastalarında geliştirilmiş olduğu görülen Amyotrofik Lateral Skleroz', 'lipopolisakkarit biyosentezi' ve 'glutatyon metabolizması' yolaklarının aktif olduğu gözlenmiştir. Bu sınıflandırmada elde edilen doğruluk %89 idi. Sonuç olarak, bakteriyel bağırsak mikrobiyom bileşimleri ve bunların potansiyel hedef yolları belirlenmiştir. MS hastalarının sınıflandırılmasında ve kontrollerde potansiyel belirteçler önerilmiştir. Mikrobiyom bileşimi ile konağın bağışıklık sistemini birbirine bağlamak için yapılacak diğer çalışmalar, hastalığın gelişiminin ve ilerlemesinin ortaya çıkmasında büyük bir etkiye sahip olabilir ve MS' e karşı yeni tedavisel stratejiler önerilmesine yardımcı olabilir.

Anahtar Kelimeler: Bağırsak Mikrobiyomu, Multiple Skleroz, Mikrobiyom Analizi, Yolak Analizi, Makine Öğrenmesi

1. BACKGROUND AND AIM OF STUDY

One of the functions of the immune system is to differentiate non-self from a self. By differentiating non-self, the immune system protects the body by responding to pathogens, such as viruses, bacteria, and parasites. Compromised by various reasons, immune cells make a mistake and attack self-cells. These mistakes can lead to a broad category of autoimmune diseases. These diseases can affect one, ten, one hundred or a thousand people per million. Other than true-autoimmune diseases, there are various diseases whose conditions considered to be related to autoimmune disease. One of the diseases that its conditions might be related to autoimmune disease is Multiple Sclerosis (MS). MS is a nervous system disease that affects the human brain and spinal cord via damaging the myelin sheath. Myelin sheats' primary function is protecting nerve cells by being surrounding material. The damage on the myelin sheath slows down or blocks messages between the brain and the body, leading to MS's symptoms (1–4).

In most cases, MS is mild, but seldomly people lose the ability to write, speak, or walk. There is no cure for MS, but medicines (immunomodulatory: IFN-beta and glatiramer) may help manage the disease and improve the quality of life by slowing it down and helping control symptoms. Some of the validated physical and occupational therapies help to relieve the MS symptoms. Both the physical activity and immunomodulation also affect and also affected by the microbiome. Although some good quality microbiome projects have been published in the past, the improved bacterial taxonomy through this time makes previous data reanalyzed for better resolution for microbiome& MS interactions. Therefore, this thesis aims to investigate whether the human gut microbiome alterations in MS with improved bacterial taxonomies with various feature selection and decision tree tools (5,6).

Bacterial taxonomies are still developing with a tremendous amount of sequencing facilities. Our increased amount of the whole genome sequencing and improved bacterial taxonomies help us obtaining a better microbial profile of previously analyzed data. Not only the resolved unclassified sequences via higher identifications but also the corrected bacterial taxonomies help us higher resolution in microbiome profiles. In this thesis, concerning a various number of machine learning, *Lysinibacillus, Blautia, Akkermansia*, and *Anaerostipes* levels are observed with the help of improved operating taxonomic units (OTUs). Testing datasets can be explained higher than 87% at least accuracy using various machine learning algorithm.

Other than the direct immunological effects, the indirect empirical metabolite effect of the microbiome from microbial diversity also tested in this thesis. Usually, researchers use both LS-NMR/MS and metatranscriptomic sequencing for the gut metabolites' role in the development of MS in patients. Although Mallick *et al*., 2019 performed predictive modeling of metabolome from 16S sequencing data with the help of the Human metabolome project, predictive metabolome modeling has performed for the first time with the help of the Human microbiome project and GhostKoala database (7). From that novel empirical pathway analysis, Amyotrophic lateral sclerosis, lipopolysaccharide biosynthesis, and glutathione metabolism found to be significantly important by using Gradient Boosted Random Forest classified decision trees. The accuracy of classification of a decision tree test data is at least 83%. Various researchers will highly use our method for early hypothesis for various other diseases.

With the help of the submitted thesis, managing disease with fortified therapeutics with sufficient support from microbiome modulation will be evaluated by MS researchers. Our novel results will be used for adjunctive therapies concerning microbiome modulations and ignite new project proposals. It has to be further investigated experimentally.

2. INTRODUCTION

2.1 Multiple Sclerosis

Multiple Sclerosis (MS), also known under its Latin name encephalomyelitis disseminate (ED), is common an immune-mediated and Central Nervous System (CNS) demyelinating disease. There are nearly 2,500,000 Multiple Sclerosis people in the world (4). The World Health Organization (WHO) reported that the mean worldwide predominance of the MS is assessed to be 30/100 000, with an average rate of 2.5/100 000 and a normal period of the beginning of 29.2 years. The predominance is most noteworthy in Europe (80/100 000) and more prominent in high-pay nations (89/100 000) than in upper-center (32/100 000), lower-center (10/100 000), and low-salary country (0.5/100 000). (3). Patients are grouped into three major categories based on the symptoms progress of the disease. These are Relapsing-remitting MS (RRMS), Secondary Progressive MS (SPMS) and, Primary progressive MS (PPMS).

The most common form of MS is relapsing-remitting MS. The relapsing-remitting form of MS is characterized by complete or partial remission after a distinguishable MS episode leading to exacerbation of clinical symptoms. Moreover, its early symptoms include paresthesia, visual problems, and strain-induced weakness of the legs. SPMS or PPMS, then again, are described by interminable, consistent sickness movement. Relapses are characterized as the emergence of new clinical symptoms that last longer than 24 hours, which can be distinguished from the previous section at least 30 days separated and cannot be attributed to a alter in body temperature level or viral infection (1,8).

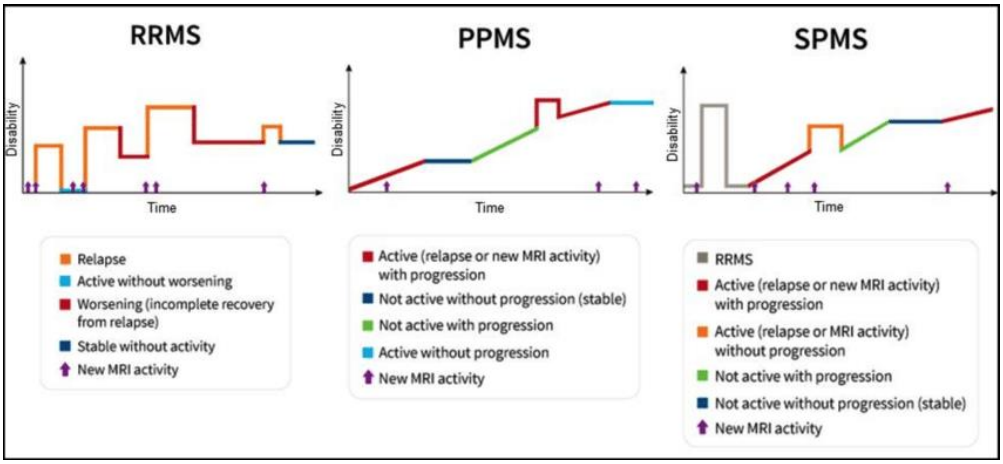

Figure 2.1. *Types of Multiple Sclerosis (Lublin et al., 2014)*

There are treatments only to reduce partially effects of symptoms, but there is no cure. The causing effect of MS is still not apparent. MS progress is influenced by both genetic and nongenetic triggers, such as a virus, metabolism, and environmental factors. The combinations of genetics and nongenetic factors cause a self-sustaining autoimmune disorder disease, MS (2,9).

2.1.1 Genotypic Effects

It is now well established that one of the most causing effects of MS is genetic susceptibility. The International Multiple Sclerosis Genetics Consortium showed that alleles of interleukin-2 receptor α gene (IL2RA) and the interleukin-7 receptor α gene (IL7RA) were strongly associated with Multiple Sclerosis (9). In 2017, Yuan Zhou et al., conducted a genome-wide association analysis (GWAS) to recognize genetic factors that predict MS relapse risk (severity), using three unique longitudinal MS cohorts. They discovered the interaction between relapse risk and the gene lipoprotein receptor protein 2 (LRP2) (10). An animal study indicated that the neurons and oligodendrocytes expressed LRP2. LRP2 manage brain development and axonal guidance. LRP2 SNP rs12988804 is the first case of a genome-wide significant association with severity in MS (11). Another investigation of genetic susceptibility

factors study with 182 children and 141 adult MS patients' groups showed that AHI1 (rs11154801) has two copies of MS risk allele and increases relapse rate among both children and adults (12).

2.1.2 Epigenetics Effects

The epidemiological studies considered that epigenetic modifications like DNA methylation, histone modifications, and miRNA expression profile play a crucial role in the development of MS disease. Epigenetic mechanisms regulate the transcription of most MS-associated genes and perpetuation in CNS cells and immune cells (13). In 2014, using Illumina 450K methylation arrays, Graves et al. examined a genome-wide DNA methylation analysis of CD4+ T cells in 30 RRMS patients and 28 healthy groups. The study found 74 significantly different methylated CpG sites in this cohort and the significant effects of CpG in the human leukocyte antigen (HLA)-DRB1 region (14). Another study observed a shift toward histone acetylation in the white matter of the frontal lobes of aged subjects and chronic MS patients. Additionally, the high level of histone acetylation was correlated with disease duration (15). A study by Aung et al. demonstrated that while protein expressions level of MMP-9 were increased in B cells during a 21 RRMS patients' relapse, microRNA-320/320a expression was decreased. Moreover, MMP-9 protein expression and secretion of MMP-9 were inhibited by microRNA-320a (16).

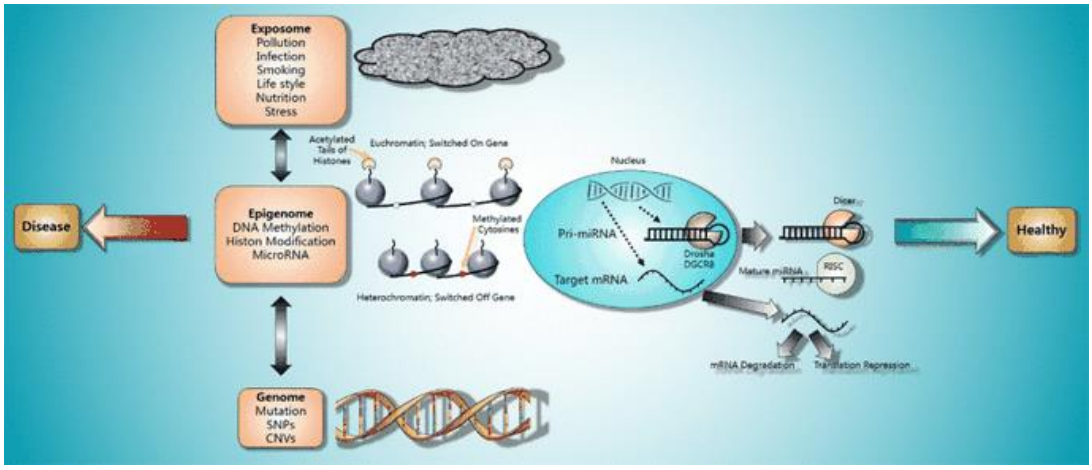

Figure 2.2. *Overview risk factor for Multiple Sclerosis (Aslani et al., 2017)*

2.1.3 Environmental Effects

Multiple Sclerosis has involved not only genetics but also lifestyle and environmental factors that may contribute to the risk of disease. These factors are Epstein–Barr virus (EBV) infection, smoking, low vitamin D, or lack of sun exposure, obesity and, gut microbiota (17). In a study, 192 MS patients and 384 health controls' cotinine levels were measured using immunoassay. They showed that A significantly increased risk of Multiple Sclerosis was associated with elevated cotinine levels (average ten ng / ml) (18). On the other hand, one of the most significant candidate viruses is Epstein-Barr virus (EBV) increases the risk of MS. In a genome-wide association study (GWAS) of anti-EBV nuclear antigen-1 (EBNA-1) IgG titers in 3599 individuals demonstrated that the genetic risk of high anti-EBNA1 titers was positively correlated with the development of MS (19).

The other most significant environmental effects for MS disease is gut microbiota. In a study of 31 MS patients and 36 healthy controls, researchers showed that the fecal microbiome of MS patients had different microbial community profiles in contrast to

healthy control group microbiomes. While *Pseudomonas, Mycoplasma, Haemophilus, Blautia,* and *Dorea* were increased in MS patients, the abundance of *Parabacterioides, Adlercreutsiza,* and *Prevotella* genera were elevated in healthy controls (6). Another study showed that the gastrointestinal and oral bacteria-derived lipodipeptide, Lipid 654, the expression level was significantly lower in the serum of 17 MS patients. Produced from commensal bacteria, Lipid 654 functions as a human and mouse Toll-like 2 ligand receptors (20).

2.2 Human Microbiome

The microbiome contains trillions of microorganisms, which are bacteria, fungi, parasites, archaea, and viruses. These microorganisms inhabit different human body parts, especially the gut, establishing the vast majority of the cells in the human body. Therefore, the human body has the most considerable diversity and abundance of microorganisms (21). Because microorganisms grow in distinct microenvironments, it has been difficult to culture them in laboratory conditions. Metagenomics approaches, where culture-independent sequencing is a stake, has provided insights into microbial communities. Recent work from the National Institutes of Health (NIH) has started the human microbiome project in 2017 to provide resources to scientists to gain taxonomic and functional perspectives about the complexity of human microbial communities and their potential roles in human health and diseases. The human microbiome project revealed the presence of complex microbiota at various surfaces of the human body, such as nasal passages, oral cavity, skin, gastrointestinal tract, and urogenital tract (22).

The composition of the human microbiome is special for each individual, and the variations between individuals are greater than the usual biochemical variations that occur over time within a person. Interaction of humans with the environment provides the potential for different microbial taxa to either serve as an immune stimulant. It can affect the microbiome through such as inflammation, fungi, and viruses that may

colonize the human body (23). The microbial metabolome interactions with the immune, endocrine, and nervous systems are associated with numerous diseases, including inflammatory bowel disease, multiple sclerosis, diabetes (types 1 and 2), allergies, asthma, autism, and cancer (24–26).

2.2.1 Gut-Brain axis and Microbiome

The brain communicates with the microbiota is bidirectional, by various pathways: neurological through the VN and/or spinal cord, endocrine (through hypothalamic pituitary adrenal, HPA, axis), immune (cytokines) and metabolic such as short-chain fatty acids, (SCFAs), tryptophan, etc. Bacteria release neuroactive compounds such as π-aminobutyric acid (GABA), serotonin, dopamine, acetylcholine (ACh), and act locally on the enteric nervous system, namely the gut-brain (27,28). It is well recognized that 90 % of the total body serotonin is synthesized in the intestine and has a significant effect on the physiology of the gastrointestinal tract; it works as a neurotransmitter. Our diet is important in this sense since this serotonin is produced from tryptophan, an essential amino acid that is only obtained from food (29). L-Glutamate (Glu) and aminobutyric acid (GABA) are known primarily for their role in mammalian central nervous system (CNS) as the main neurotransmitters with thrilling and inhibitory roles (30). Glu can be synthesized by several bacteria such as *Corynebacterium glutamicum*, *Brevibacterium lactofermentum*, *Brevibacterium flavum*, *Lactobacillus Plantarum*, *Lactobacillus paracasei*, and *Lactococcus* (31,32). Multiple forms of Glu receptors have been found in the stomach, small intestine, and gastrointestinal epithelial (GI) cells and/or enteric neurons. And mGlu4 receptors were detected in both the gastric antrum and duodenum mucosa, while mGlu4 and mGlu7 receptors were identified in the epithelium of the colon. High levels of mGlu7 and mGlu8 were detected in myenteric neurons, where they may be involved in gut motility regulation (33).

Moreover, Marine microorganisms, *E. coli*, *Pseudomonas*, several LAB (e.g., strains belonging to *Lactobacillus*, *Lactococcus*, and *Streptococcus* genera) and *Bifidobacterium* strains are able to produce GABA (34,35). The GABAB receptors are expressed widely in GI tractors. GABA and its ionotropic and metabotropic receptors are widely distributed throughout the enteric nervous system (ENS) from the abdomen to ileum, in both submucosal and myenteric neurons. 5-HT released by endothelial cells in the small intestine of the guinea pigs is modulated by GABAA and GABAB receptors. GABAB receptors have been reported to be involved in modulating vagal and spinal sensitivity (33). Besides, microbial metabolites such as propionic acid, butyric acid, and acetic acid can induce neuromodulatory substances to be secreted by epithelial enterochromaffin cells, neurons, or immune cells. For example, butyric acid has anti-inflammatory and neuroprotective properties through histone deacetylases inhibition (36).

2.2.2 Disease and Microbiome Alteration

The gut microbiota's impact on human health is continuous from birth to old age. The microbiome is composed of microbes, which can be both beneficial and potentially harmful. Most are symbiotic, and some are pathogenic, which promote illness in smaller numbers. In a healthy human, there is a balance between pathogenic and symbiotic microbiota without any problem. However, due to the usage or exposure to antibiotics, specific diets, or irregular sleeping habits and food evolution may cause a disturbance in that balance, which causes dysbiosis. Appropriately, the diverges from healthy microbiome has been related to many diseases and disorders, such as metabolic disorders, neurodegenerative, and immune-related diseases (37–40).

2.2.2.1 Metabolic Disorder

Microbiome researchers have already described interactions in various human metabolic diseases. In a study of 20 obese subjects, nine patients with anorexia nervosa versus 20 normal-weight healthy controls, while *Bacteroidetes* community was reduced in obese patients, the abundance of *Lactobacillus* species was increased in obese patients than in lean controls or anorexic patients (38). Animal studies showed possible causal function for intestinal microbiota in non-alcoholic fatty liver disease (NAFLD). In a study, when scientist compared to 43 NAFLD patient with 83 healthy controls, they observed that NAFLD patient has lower diversity richness and different a phylum-level in their fecal microbiome contrast to controls groups (41).

2.2.2.2 Immune Related Disorder

As mentioned above, the intestinal tract of mammals contains trillions of bacteria co-evolved with the host in a symbiotic relationship. The symbiotic bacteria provide the host with many functions that promote immune homeostasis, immune responses, and pathogen colonization defense. Moreover, throughout the presence of gut microbiota, pathogens have developed strategies to facilitate replication. Pathogen infection encourages overgrowth of harmful pathobionts and inflammatory disease development (42). A study about the impact of the microbiota on the human immune system shows that polysaccharide A (PSA) producing organisms have protective associations with inflammatory conditions. As an example, the human symbiont *Bacteroides fragilis* has been shown to correct systemic T-cell deficiencies, T-helper cell imbalances, and direct lymphoid organogenesis in germ-free mice (39).

The most abundant resident immune cells in the brain are microglia. Microglia execute canonical functions of myeloid cells such as phagocytosis, antigen presentation, cytokine production, and inflammatory response activation. Microbiota affects the maturation and function of microglia. Erny D. *et al.* exhibit that mice reared in the absence of microbial colonization (germ-free, GF) have microglia with irregular morphology, altered gene expression, and impaired functional response to stimulation compared to conventionally colonized (specific pathogen-free, SPF) controls (43).

2.2.2.3 Neurogenerative Disease

Human gut microbiome influences Central Nerves System (CNS) through signaling pathways. It can trigger some diseases including dementia, Huntington's disease and Alzheimer's disease (AD), Parkinson's disease (PD), non-immune-mediated diseases including autism, depression, anxiety and stress, and an immune-mediated disease which is Multiple Sclerosis (MS) (40). Neuroinflammation is one of the primary mechanisms that link the microbiota to age-related diseases. The gut microbiota plays a crucial role in microglia activation, and it has been suggested that the gut microbiome manipulation, specifically with bacteria producing short-chain fatty acid (SCFA), may modulate neuroimmune activation (44).

In similar finding as previous study of 25 AD patients and 94 healthy controls, researchers showed that the fecal microbiome of AD patients had different microbial community profiles in contrast to healthy control. While *Firmicutes* and *Bifidobacterium* were decreased in AD patients' groups, *Bacteroidetes* were increased in the microbiome of AD participants (45). Another acute, chronic CNS demyelinating disease is MS intermediated by an auto-reactive immune attack against central neural tissues. In a study including 31 MS patients and 36 healthy controls, researchers showed that the fecal microbiome of MS patients had different microbial community profiles compared to healthy control group microbiomes. While *Pseudomonas*,

Mycoplasma, Haemophilus, Blautia, and *Dorea* were increased in MS patients, the abundance of *Parabacterioides, Adlercreutsiza*, and *Prevotella* genera were elevated in healthy controls (6).

2.3 Bacterial Diversity and Taxonomic Classification

2.3.1 Metabolic Diversity of Microorganisms

A healthy human gut microbiota covers a wide range of representatives of bacterial species, both phylogenetically and metabolically diverse. These representatives play a crucial role in carbon, nitrate, lipids, and energy metabolism.

2.3.1.1 Carbon Metabolism

At any given time, there are many different forms of carbohydrates present in the large intestine, with the concentrations of each substrate continually changing as they are broken down, replenished, or replaced. Catabolite control mechanisms, containing catabolite suppression, catabolite inhibition, and inducer exclusion, regulate the carbohydrate transport in many prokaryotes. The intestinal bacteria produce the primary short-chain fatty acid (SCFA) during the breakdown of carbohydrates, including butyrate, acetate, ethanol, lactate, succinate, and propionate (46).

Recent studies show that lower abundances level of butyrate-producing bacteria *Bifidobacterium* species in the human colon has belonged to antibiotic-associated diarrhea, IBS, IBD, obesity, allergies, and regressive autism (47). Qiu *et al.* showed that the butyrate producer *F. Prausnitzii* produces anti-inflammatory

peptides that block the activation of the NF-πB nuclear factor and cytokine development IL-8 in mice and provide protection against chemical colitis (48). *Veillonella* makes use of lactate as its sole source of carbon. *Veillonella* atypica enhances run time through its metabolic conversion from exercise-induced lactate to propionate, thereby recognizing a normal, microbiome-encoded enzymatic mechanism that enhances athletic efficiency (49)

2.3.1.2 Nitrate Metabolism

Recent findings into the inorganic bioactivation and signaling actions, dietary nitrate and nitrite presently advise a crucial role for the microbiome in the development of some disease. Research in the 1970s demonstrated that commensal bacteria could be involved in the pathogenesis of gastric cancers and other malignancies by reducing nitrate to nitrite since nitrite can enhance the carcinogenic N-nitrosamines generation. On the other hand, other studies suggested that Nitrate-to-nitrite bacterial metabolism and the subsequent production of biologically active nitrogen oxides can be beneficial (50). The nitrate bioactivation from dietary or endogenous sources such as oxidation of NO needs its initial nitrite reduction, and since mammals have no specific and effective nitrate reductase enzymes, this transformation is done mainly by commensal bacteria (51). *Granulicatella adiacens, Haemophilus parainfluenzae, Actinomyces odontolyticus, Actinomyces viscosus, Actinomyces oris, Neisseria flavescens, Neisseria mucosa, Neisseria sicca, Neisseria subflava, Prevotella melaninogenica, Prevotella salivae, Veillonella dispar, Veillonella parvula,* and *Veillonella atypica* are most known nitrate-reducing bacteria in the oral cavity by identified through 16S rRNA gene pyrosequencing and analysis (52).

Studies showed that nitrite produced from the reduction of bacterial nitrate is a significant NO storage pool in blood and tissues where NOS-mediated development is insufficient so, dietary nitrite and nitrate might protect the heart from heart attack

injury (53,54). A study from 1991s showed that gastric nitrogen oxides other than N-nitrosamines protect the gastric mucosa and inhibit *Helicobacter pylori* growth as a causative factor in the development of gastric cancer in humans (55).

2.3.1.3 Lipid Metabolism

The blood lipid levels monitoring is frequently used in the clinic to assess disease risk and prevent the development of the disease. In a study of 145 European women with normal, impaired, or diabetic glucose control, scientists analyzed that there is a negative correlation between *Clostridium* and serum triglyceride levels and a positive correlation between high-density lipoprotein (HDL) and *Clostridium* species (56). A study by Le Chatelier *et al.* demonstrated that 169 obese Danish individuals had lower bacterial richness characterized by a higher level of insulin resistance, triglycerides, free fatty acids, and overall adiposity, whereas had a lower level of HDL- cholesterol and inflammatory phenotype than 123 nonobese Danish groups (57). Gut microbes also affect host physiology by modifying host-synthesized bile acids (BAs). BAs function as hormones by their ability to activate hormone receptors and G-coupled protein receptors. They modulate homeostasis of glucose, lipid metabolism, energy expenditure, and motility of the intestines (58). Type 2 diabetes (T2D) subjects have altered the BA profiles in circulation. Treatment of T2D subjects with compounds that enhance Bas fecal excretion and change composition of BA improves their glycemic status (59).

2.3.1.4 Energy Metabolism

The gut microbiota evolved as an environmental factor that modulates the energy balance of the host. It enhances the host's capacity to extract energy from the digested food and generates metabolites and microbial products, including short-chain fatty acids, secondary bile acids, and lipopolysaccharides. These metabolites and microbial

products serve as signaling molecules that regulate appetite, gut motility, storage, and energy consumption (60). In a study, 29 overweight and 41obese pregnant women fecal microbiota profiles revealed that the levels of metabolic hormones and microbiome profiles differentiated between overweight and obese women. The rates of adipokine which is fasting metabolic hormone were closely associated with *Ruminococcaceae* and *Lachnospiraceae,* which are dominant energy metabolism families. Insulin related positively to the *Collinsella* genus. The gastrointestinal polypeptide was significantly associated with the *Coprococcus* genus but adverse to the *Ruminococcaceae* family (61).

2.3.2 Bacterial Taxonomic Classifications

Advances in high-throughput DNA sequencing and analyzes of bioinformatics have illustrated the essential role of microbial communities in human populations and planetary health and facilitate massive microbiome meta-analysis (62,63). There are two main approaches to microbiome analysis, 16S ribosomal RNA (rRNA) gene amplicons, and shotgun metagenomics, which are illustrated by library analyses designed to highlight their strengths and weaknesses. Classification of short marker-gene DNA sequences such as bacterial 16S rRNA genes to infer taxonomic composition is crucial in characterizing microbial communities (64). 16S rRNA regions are approximately 1600 base pairs long, and they are the most significant factor for the identification of community composition. There are three main reasons why scientist prefers 16s rRNA to the classification of bacteria. Firstly, Its existence in almost all bacteria, often as a multigene family, or operons. Secondly, the 16S rRNA gene function has not changed over time, indicating that random sequence changes are a more reliable time measure. Lastly, the 16S rRNA gene is big enough for use in computer science (65). Although the sequencing cost of extended read length is pricy, it can span multiple hypervariable regions of the 16S rRNA gene, which then can classify at a species level. The short-read sequencing platforms have many PCR primers to amplify 16S rRNA regions of different hypervariable regions for sequencing (66). 16S rRNA full-length gene sequences consist of nine hypervariable

regions divided by nine strongly conserved regions. These are V1, V2, V3, V4, V5, V6, V7, V8, and V9. The 16S rRNA gene's methods depend on polymerase chain reaction (PCR) using universal primers aimed at preserved regions to amplify as wide as possible many different microorganisms (67).

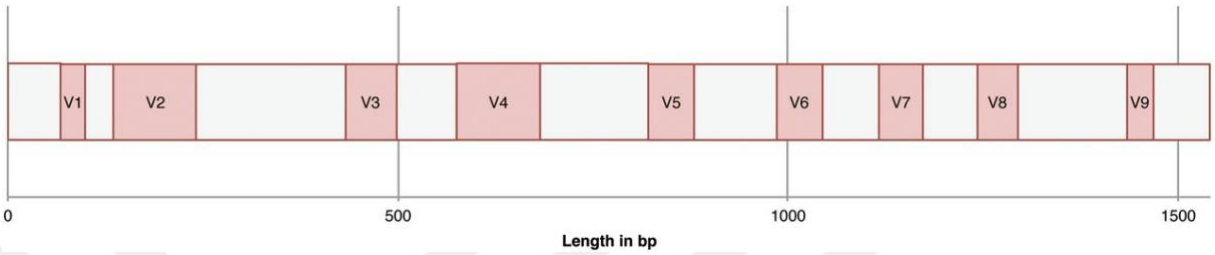

Figure 2.3 *Approximately 1.5 kb 16S rRNA E. coli gene that displays the nine variable regions*

(Hum Mol Genet, Volume 22, Issue R1, 15 October 2013, Pages R88–R94, https://doi.org/10.1093/hmg/ddt398)

Typically, the procedure for 16S analysis involves consistency filtration, error correction, elimination of chimeric sequences, clustering of reads into 'Operational Taxonomic Units' (OTUs) based on sequence similarity and classification of OTUs. A popular first step is to run a variety of quality control computational tools which identify and remove sequences and contaminants of low quality such as FastQC, Cutadapt, BBDuk and Trimmomatic. PCR-amplified 16S sequences were clustered based on similarity to the generation of operation taxonomic units (OTUs) and representative OTU sequences compared with the reference databases to indicate likely taxonomy. When using partial 16S rRNA gene sequences, a rate of 97 percent sequence identity is frequently selected as representative of a species and 95% for a genus (68). Identification accuracy relies on the selected reference database, such as The Ribosomal Database Project, GreenGenes, and SILVA (69). Genome sequencing is one of the most powerful tools available in microorganisms to access genetic and metabolic diversity. The sequencing stage produces millions of sequences. Such sequences, known as 'reads,' can be considered to reflect the compositional properties of the genomes of their source. Analyzing these readings can thus provide insights into the composition of various microbes that constitute a microbiome. As of July 2013,

the microbial genome site of the National Center for Biotechnology Information (NCBI) listed 2552 whole genomes. It is now possible to map all the genes which characterize a particular group of organisms by sequencing a wide range of isolates from various sources (70).

2.3.2.1 K-mer Based Classification

With 16S sequences, the most common pattern recognition methods are those focused on counting k-mers, i.e., overlapping 'words' of length K in the sequences. The ribosomal database project (RDP) classifier is based on the naïve Bayes principle and a word-length of K=8 and is now near-standard in 16S based classification. The RDP classifier uses the K-mer count in one of several different ways (71). The methods of K-mer are fast and will not suffer from the same uncertainties as evolutionary models and alignment procedures. This method of translating sequences into numerical data is not as intuitive as evolutionary models and lacks the clear definition of evolutionary distances, but in their process, they are rather objective. The word k-mer applies to all subsequences of length k, so that the AGAT series will have four monomers (A, G, A, and T), three 2-mers (AG, GA, AT), two 3-mers (AGA and GAT) and one 4-mer (AGAT), respectively. (72). There are several metagenomics tools for classification of metagenome datasets such as CompostBin ($k = 6$), PCAHIER, PhyloPythia ($5 \leq k \leq 6$), CLARK ($k \geq 20$), TACOA ($2 \leq k \leq 6$) and KRAKEN2 (73–78). KRAKEN was the first way to easily classify all reads in a metagenomic sample. KRAKEN creates a database that stores the species identifier for that k-mer, with each k-mer in each genome. Database k-mers and their taxa are stored in a compressed lookup table, which can be quickly searched for exact matches to k-mers found in the metagenomics dataset reads. Its default parameter is k=31, but the database can be constructed with any length k-mer. The k selection represents a significant trade-off between sensitivity and specificity: excessively long k-mers can fail to match due to sequencing errors or genuine differences between species and strains, whereas excessively short k-mers give many genomes a false match (79).

2.3.2.2 Alignment Based Classification

Metagenomic classification tools align sequences against a microbial genomes database to classify the taxon of each gene. rRNA gene sequences have been the most commonly utilized marker sequences; this contained the 16S rRNA gene for bacteria, the 18S rRNA gene for eukaryotes, and the internal transcribed spacer (ITS) regions of the fungal ribosome for fungi. Sequenced genomes have been used as references in metagenome studies with known taxonomies for reading recruitment. These markers perform well for phylogenetic profiling as they are ubiquitously found in the population, they have hypervariable regions that differentiate organisms and they are flanked by preserved regions that can be targeted by 'universal' primers (80).

Alignment and mapping methods are widely used in the alignment-based approaches to locate similarity hits from the query read to the references. For example, BLAST, BLAT, and BWA are the most common tools for alignment (81). BLAST (basic local alignment search tool) is an algorithm and software in order to compare each read with all sequences, such as protein amino acid sequences or DNA and RNA nucleotide sequences with a library or sequence database and identifying library sequences that resemble query sequences above a given threshold. Summarily, by finding short matches between the two sequences, BLAST identifies similar sequences (82).

2.4 Machine Learning Algorithms

Metagenomics recognizes the need to develop computational methods that make it possible to understand the genetic composition and activities of species communities so complex that they can only be sampled, never wholly characterized. Machine learning currently provides some of the most promising methods for creating

predictive models for biological data classification (83). This algorithm is said to 'learn' if experience increases its efficiency for a particular job. Initially the model suits on a training dataset, which is a collection of examples used to match the model's parameters such as weights of neuronal connections in artificial neural networks. The model is trained on the training dataset using a supervised method of learning, for example using methods of optimization such as gradient descent or stochastic gradient descent. Throughout the practice, the training data set always consists of input vector pairs and the corresponding output vector, in which the response key is usually referred to as the target. The model parameters are adjusted based on the result of the comparison and the particular learning algorithm used (84–86).

2.4.1 Decision Tree

Decision Tree is a type of supervised learning algorithm and is one of the most popular algorithms of machine learning used in both classification and regression problems. It is a decision support tool that uses a tree-like diagram or decision model and its possible consequences, such as the results of chance events, resource costs and utility. The decision tree defines the most significant attribute and its importance, which gives best homogeneous population sets (87). Decision trees can use many measurements to determine to split a node into two or more subnodes. Selection of the algorithms is also based on goal form variables. The four measurements most widely used in the decision-tree are. These are Gini index, Gain information, Chi-Square and Reduction in Variance. The Gini Index is calculated by subtracting from one the sum of each class's squared probabilities. It favors partitions larger than this. Information Gain multiplies the probability of class times the probability of the log (base=2) of that class. Information Gain favors smaller partitions that have a lot of different values. In the end, you have to experiment with your data and the criterion for splitting (88).

2.4.2 Random Forest

Random Forest is an ensemble learning algorithm and can be used in both classification and regression problems such as decision tree. RFs creates a large number of decision trees at the time of training and generates the class mode of the individual trees classes. (88). This subsetting and subsampling scheme is based on bagging (89). When a prediction is made later, the estimates in these decision trees are averaged. It is widely used in many domains and on many tasks successfully. Although, they have generally better performance than a decision tree, they are less interpretable (90). The method uses quantitative microbiome profiles including relative abundances at the species level, and the presence of strain-specific markers as tools. The structure is completely auto, such as selection of model and feature allowing for a systematic and non-overfit analysis of large metagenomic datasets. Cross-validation produced high disease prediction capabilities which were generally enhanced by the selection of features and the use of strain-specific markers instead of taxonomic abundance at the species level (91).

3. MATERIALS AND METHODS

3.1 Metagenomic Datasets

In this study, we have used two different SRA datasets. In the first dataset, samples of 16S rRNA V4 region sequences were collected from mainly human stool (RRMS=29, Healthy=42), which were produced by Illumina MiSeq with 150-bp paired-end (PE) reads submitted by Brigham and Women's Hospital. For the second dataset, samples of V3-V5 regions of the 16S rRNA stool sequences were obtained from human stool (RRMS=31, Healthy=36), which were sequenced by Illumina HiSeq 2000 and submitted by Mayo Clinic. They were downloaded from the Sequence Read Archive (SRA) of the National Center for Biotechnology Information (NCBI) with accession no PRJNA321051 and PRJNA335855. Raw reads were processed using the BBMAP tools software package (v.38.61), which performed quality filtering and adapter trimming.

3.2 Taxonomic Analyses and Cluster Quality

Operational taxonomic unit (OTU) tables were constructed with KRAKEN2 software (v.2.0.8) with RefSeq (bacteria) database as of December 2019 at the 99% identity level (0.01 confidence). The taxonomic assignment of the sequences also was performed by using BLAST-N (v.2.10, dbV5) against the NCBI NT database with a cut-off 90% identity. KRONA (v.2.7.1) was used to visualization and complex hierarchies of metagenomic classifications.

3.3 Predictive Modeling using Random Forests

We have used Random Forests (RF) to predict the disease status based on the microbiota profile (genus-level relative abundance data) using default parameters of the R implementation of the algorithm (R package "randomForest"). Bootstrapping (n = 500) was used to assess classification accuracy. The classification performance was compared to random guess, where the class label of a test sample was predicted to be the label of the majority class in the training set, and the significance of difference was assessed by an Independent student t-test. Decision Tree was used to predict to make accurate predictions disease based on the genus and species level microbiota profile using default parameters of the R implementation of the algorithm (R package "rpart"). All the statistical analyses were performed in R-3.0.2 (R Development Core Teams). Decision tree models are used to solve classification problems as their goal is to train a decision tree on existing data and use it to classify new data.

In this thesis, workflows have prepared where the data split it into a training and a test set. The data set has numerical columns representing the measures of bacterial concentrations and one nominal column: class (MS or Healthy). The goal is to train a decision tree to classify whether a class is MS or Healthy. The Decision tree algorithm performs a split on the best input feature in each iteration, and in our case, it is the bacterial concentrations. The thesis's quality measure for choosing the most informative feature in gain ratio. A split on a numeric attribute is always a binary split, which partitions the dataset into two subsets. The Caret library for the decision tree has enough parameters for selecting an attribute from a very important or a very reliable analysis. As we have a large variety of elements in microbiome analysis, we have less risk of not learning enough than the risk of overfitting. So, the minimum split sizes should be determined concerning the number of taxonomic units existed in OTU tables. If we have a full tree with many nodes splitting just a few records, the risk of over-fitting is quite high. The number of records per node is indeed a stopping criterion for the caret algorithm. As soon as the number of records in a node is smaller than the

predefined number, the algorithm prevents further splitting on this branch. So basically, the number of the elements in each OTU table is carefully inspected for analysis. If the number of taxonomic units and person to person variations are high, in order to get a higher accurate and reliable decision tree using a higher taxonomic clade would be a good option.

3.4 Functional characterization

The Human Microbiome Project datasets which consist of human microbiome bacterial genomes have been analyzed. GLIMMER was then used to extract protein from these data sets of bacteria, and it was purified according to the first data analysis of the bacteria. Bacterial protein data is standardized based on abundance of bacteria following filtration. GoastKoala was used to predict the abundance of functional categories such as clusters of the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways based on the 16S rRNA sequence data. 291 KEGG pathways were compared and 39 differentially abundant KEGG pathways were identified which show a diverse change in the contrast between RRMS microbiota and healthy controls.

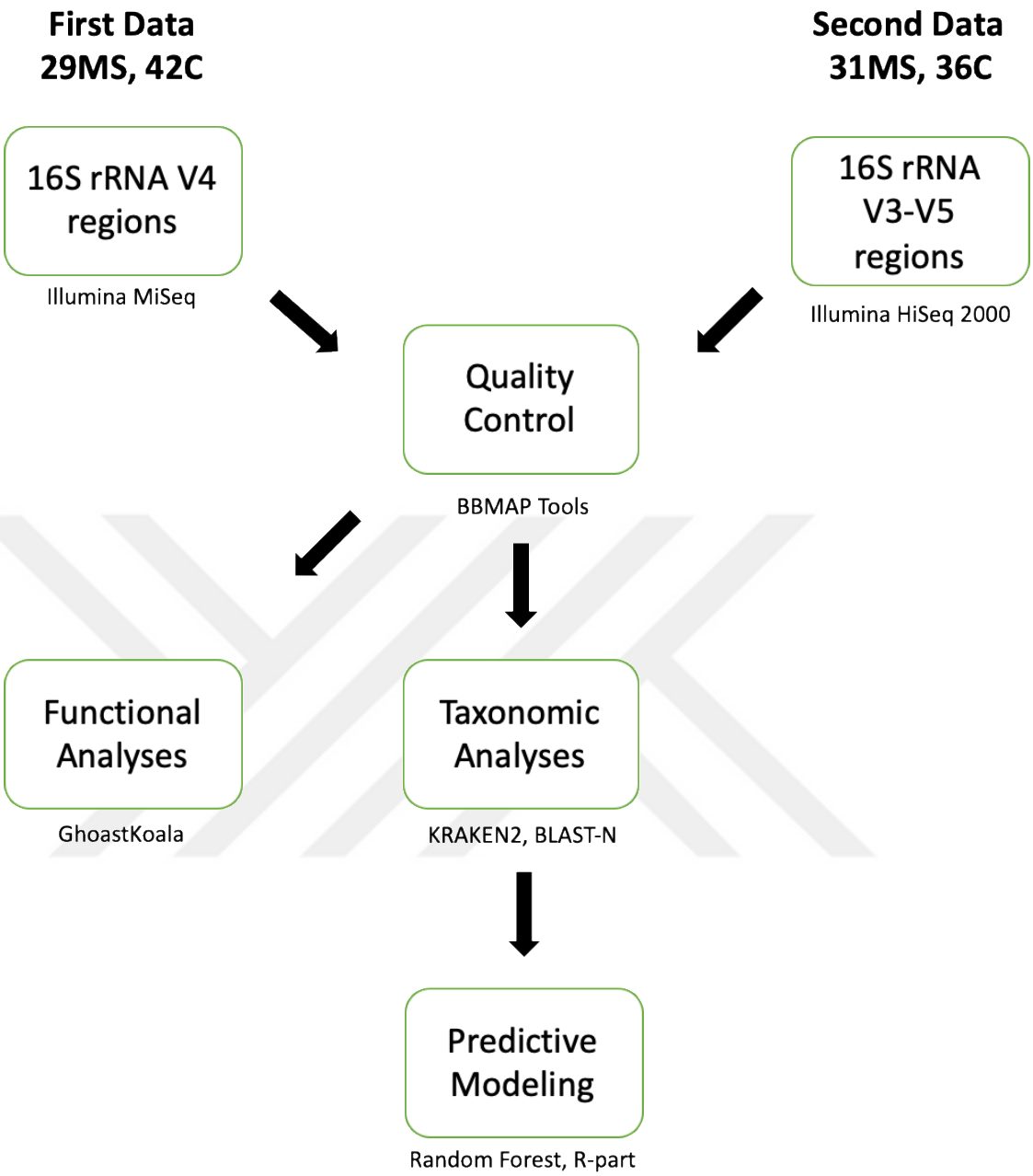

Figure 3. *Overview of the method*

4. RESULTS

4.1 Microbial Profiling of RRMS Changes at The Species Level

In the first study, 29 RRMS patients' fecal sample V4 regions sequences were analyzed using KRAKEN2 and BLAST-N. After removing adapter-leftover with using BBMAP, these sequences were clustered into OTUs based on 99% sequence similarity with KRAKEN2. 2184 OTUs were observed at the species level. *Bacteroides, Paraprevotella xylaniphila, Alkaliphilus metalliredigens,* and *Caproiciproducens sp. NJN50* showed lower abundance, whereas *Faecalitalea cylindroides* and *Ruthenibacterium lactatiformans* showed higher abundance in RRMS patients compared to healthy controls. The obtained accuracy in this classification was %71. We also perform BLAST-N to evaluate KRAKEN2 results. In BLAST-N results with cut-off 90% identity indicated that *Akkermansia muciniphila, Eubacterium ventriosum, Clostridium celerecrescens, Granulicatella adiacens,* and *Alicyclobacillus kakegawensis* showed higher abundance in RRMS patients with at least 64% accuracy.

In the second study, 31 RRMS patients' fecal sample and 36 healthy controls V3-V5 sequences adapters were trimmed with BBMAP and were Blasted with cut-off 90% identity showed that *Blautia wexlera* is the most differently abundant species. Testing datasets can be explained by 92% at least accuracy using decision tree algorithm.

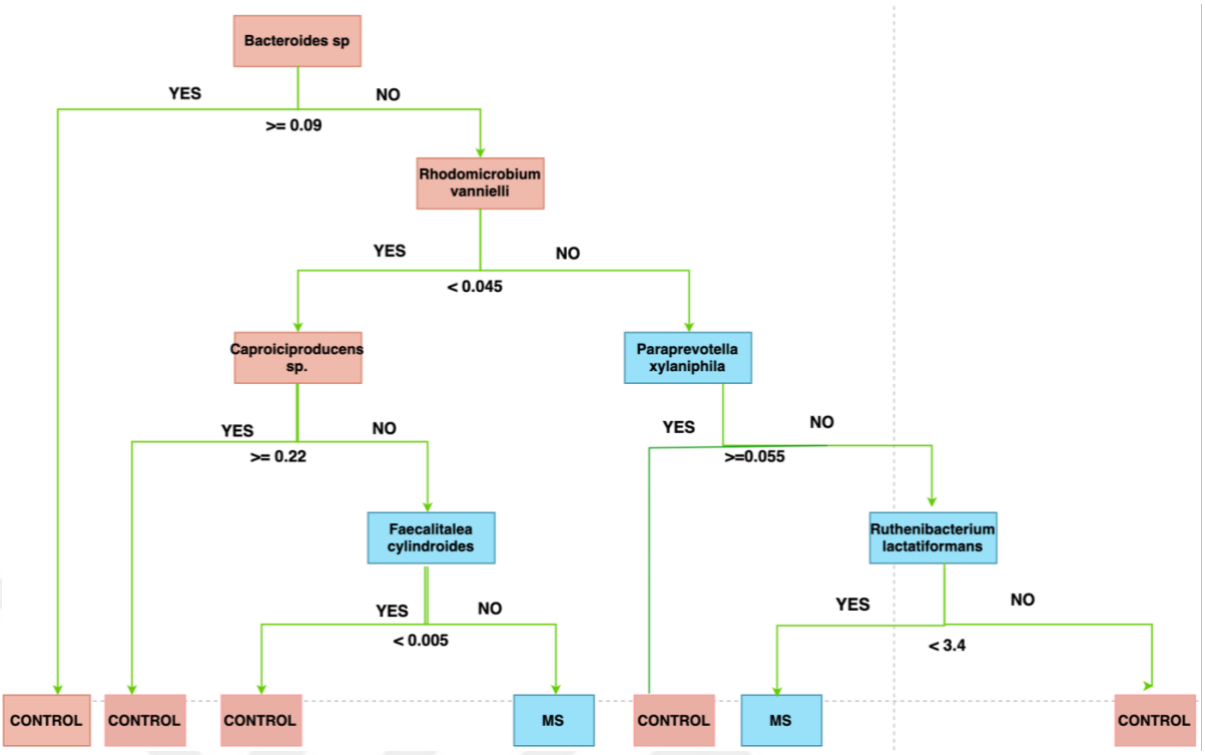

Figure 4.1. *Gut microbiota of MS patients differs from control groups. Decision Tree for first data in species-level at 71% accuracy with KRAKEN2 result. C: Control, MS: Multiple Sclerosis percentages in overall concentration.*

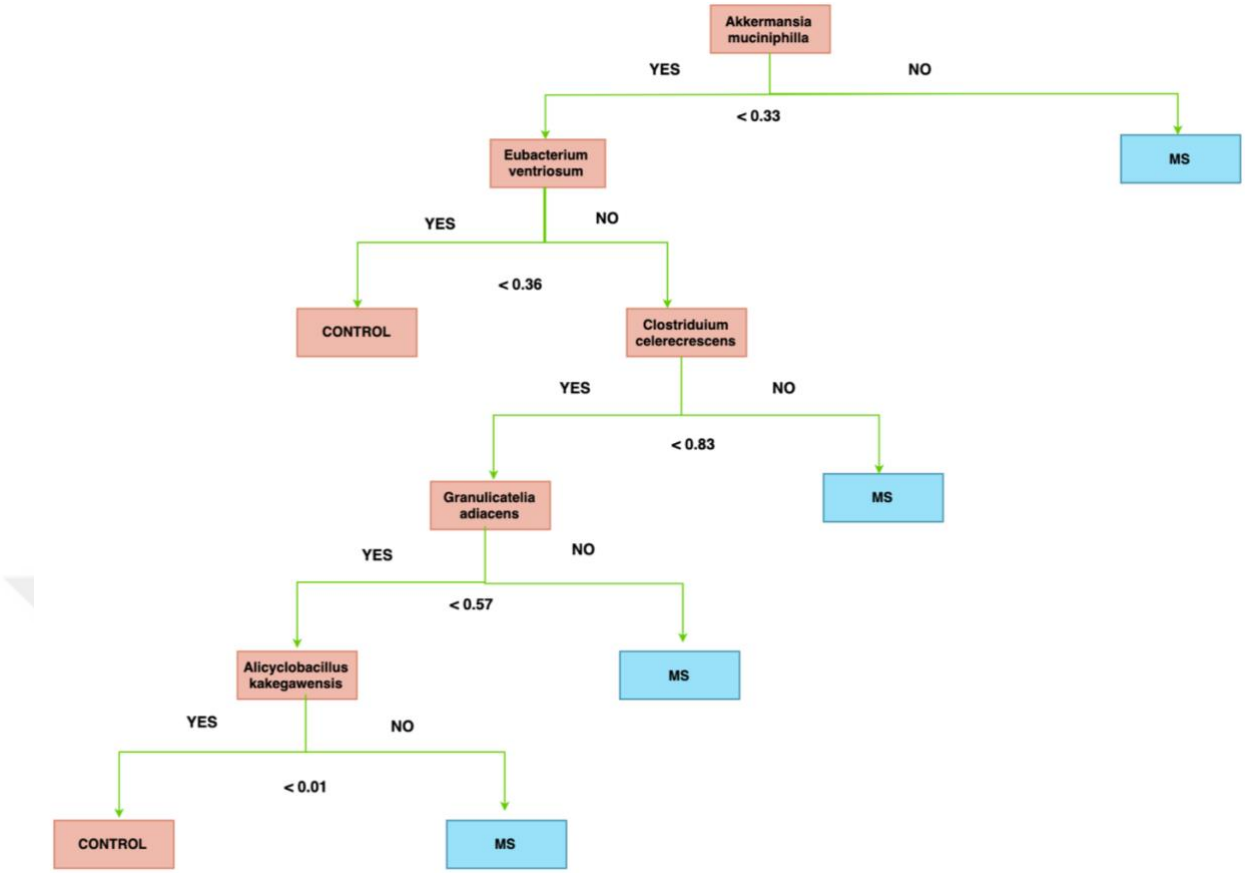

Figure 4.2. *Gut microbiota of MS patients differs from control groups. Decision Tree for first data in species-level at 64% accuracy with BLASTn result*

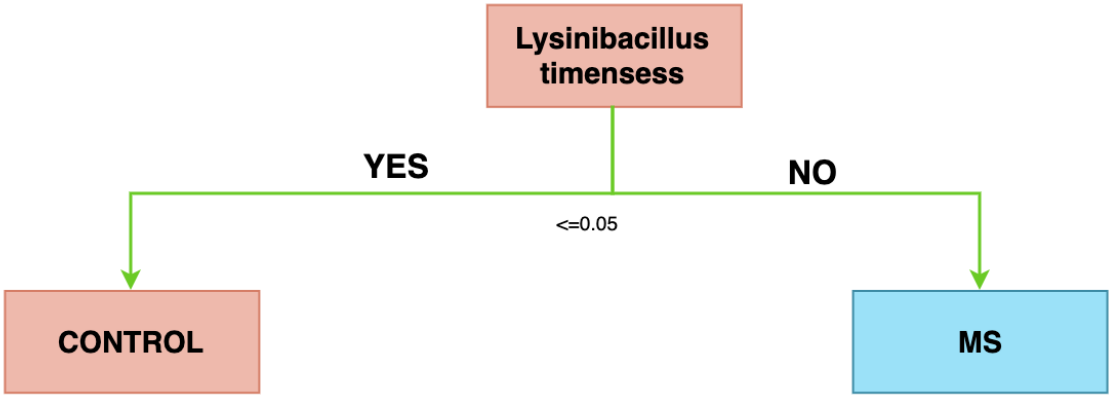

Figure 4.3. *Decision Tree for second data in species-level at 71% accuracy with KRAKEN2 result. C: Control, MS: Multiple Sclerosis percentages in overall concentration.*

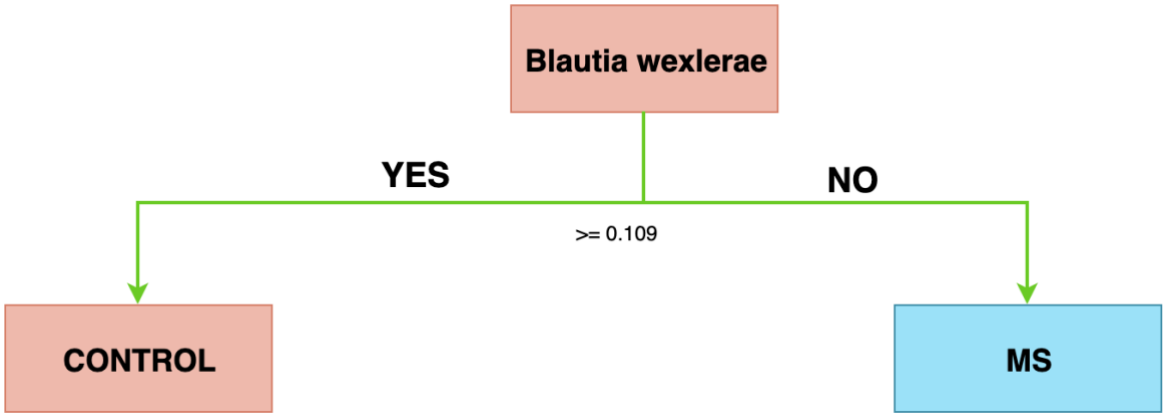

Figure 4.4. *Decision Tree for second data in species-level at 91.6% accuracy with BLAST-N result. C: Control, MS: Multiple Sclerosis percentages in overall concentration.*

4.2 Microbial Profiling of RRMS Changes at The Genus Level

In the first study, sequences were clustered into OTUs based on 99% sequence similarity with KRAKEN2 at the genus level. *Akkermansia*, *Anaerostipes*, *Mogibacterium*, and *Streptococcus* showed higher abundance, while *Bacillus* and *Anaerotignum* showed lower abundance in RRMS patients. The obtained accuracy in this classification was 65%.

In the second study, sequences also were clustered into OTUs based on 99% sequence similarity with KRAKEN2 at the genus level. While *Lysinibacillus* and *Blautia* have a higher abundance, *Butyrivibrio* is a lower abundance in RRMS patients to compared controls. The testing accuracy was 88%. In BLAST-N results with cut-off, 90% identity indicated that *Blautia*, *Dethiosulfovibrio*, and *Enterobacter* increased, whereas *Slackia* decreased in RRMS patients. Therefore, these data show that RRMS patients' gut microbiome represented differential abundance compared to control groups.

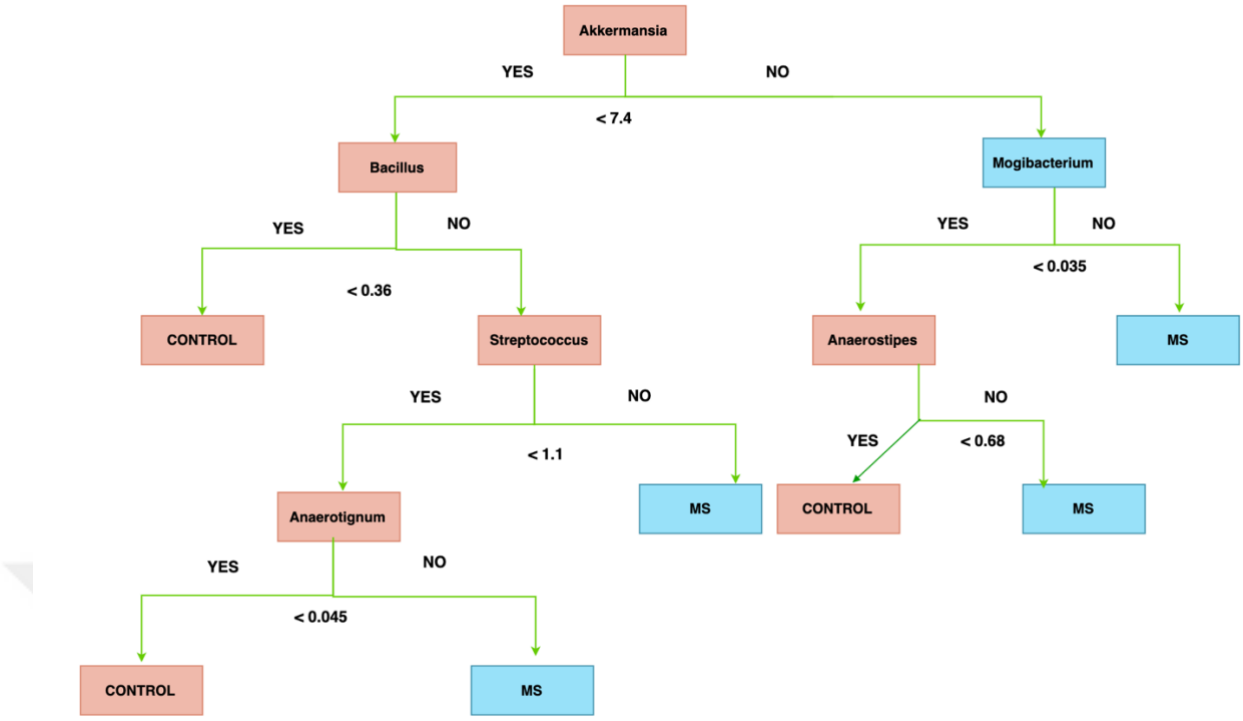

Figure 4.5. *Decision Tree for first data in genus-level at 65% accuracy with KRAKEN2 result. C: Control, MS: Multiple Sclerosis percentages in overall concentration.*

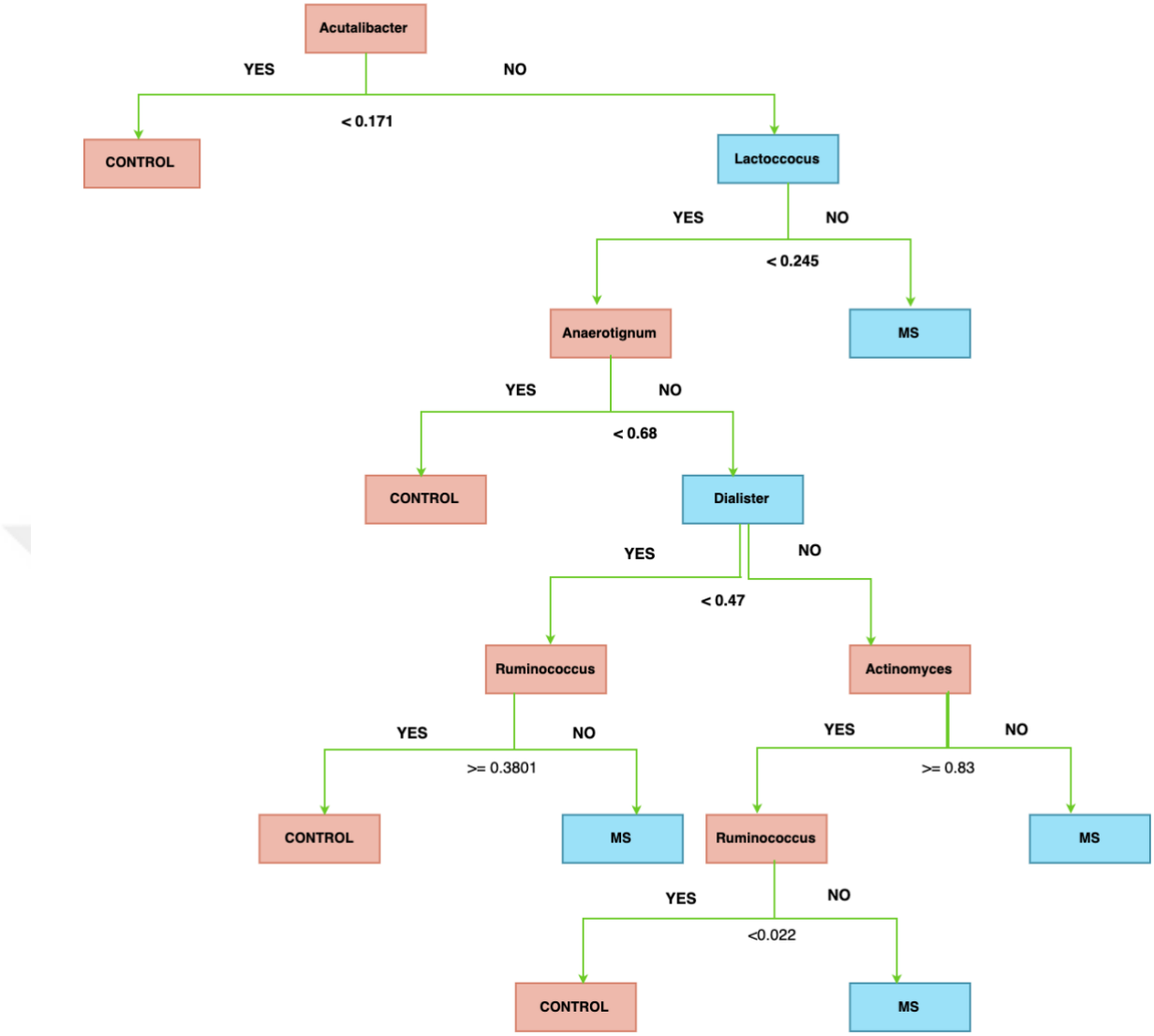

Figure 4.6. *Decision Tree for first data in genus-level at 48% accuracy with BLASTn result. C: Control, MS: Multiple Sclerosis percentages in overall concentration.*

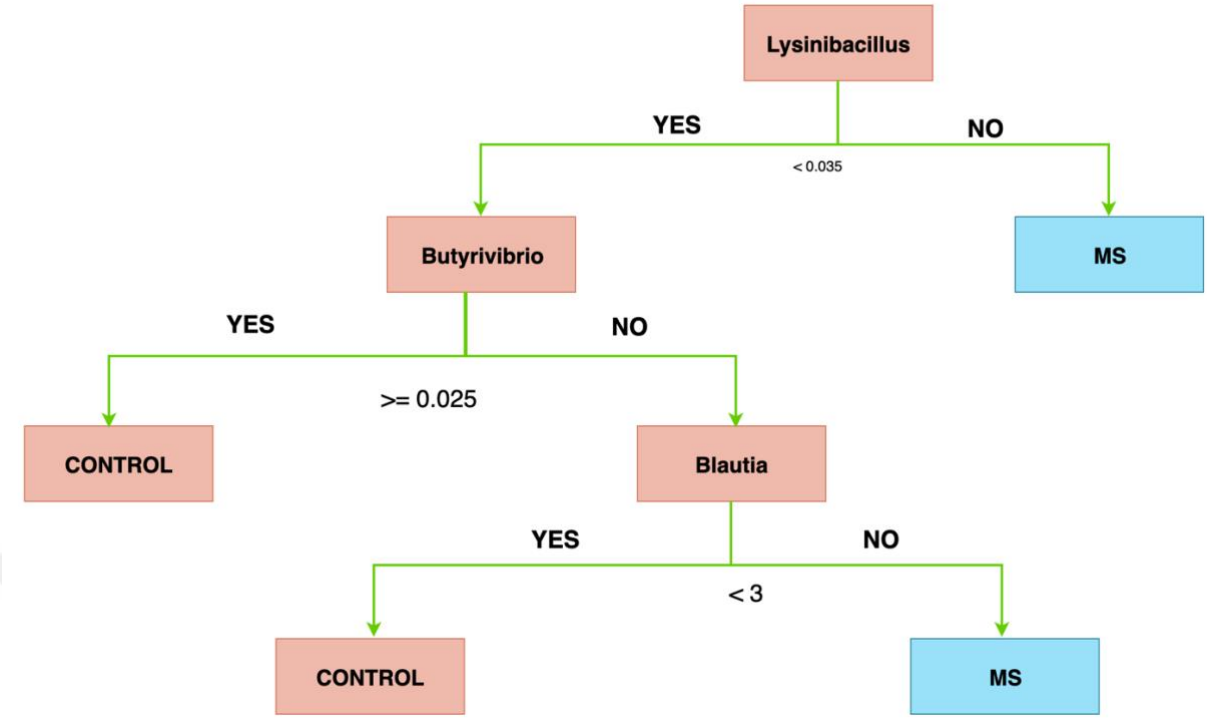

Figure 4.7. *Decision Tree for second data in genus-level at 88% accuracy with KRAKEN2 result. C: Control, MS: Multiple Sclerosis percentages in overall concentration.*

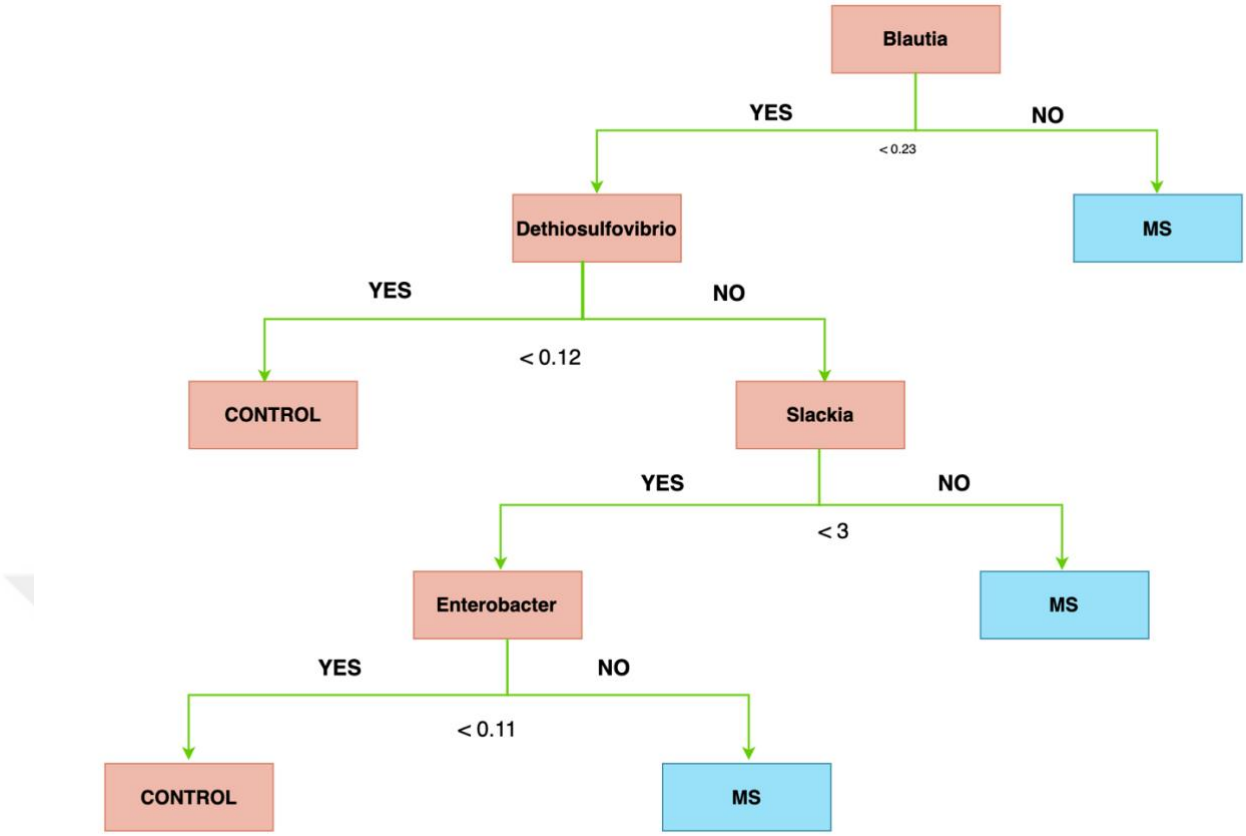

***Figure 4.8**. Decision Tree for second data in genus-level at 64% accuracy with BLASTn result. C: Control, MS: Multiple Sclerosis percentages in overall concentration.*

4.3 Random Forest

We also assessed the predictive power of the gut microbiota using Random Forest that predicts disease status based on an ensemble of decision trees. The relative importance of each genus and species in the predictive model was assessed using mean decreasing accuracy and Gini coefficient. In the first study, we just analyze at the genus level because, in species, we cannot predict with only the V4 region. *Akkermansia, Bacillus, Monoglubus, Nonomuraea, Candidatus Solibacter, Faecalitalea, Parabacteroides, Limnochorda, Hungetalla, Butyricimonas,* and *Slackia* were frequently important genus for a predictive model for KRAKEN2 results at 60% accuracy.

For the second study, *Lysinibacillus, Oblitimonas, Enterococcus, Parabacteroides, Adlercreutzia, Gordonibacter, Anaerostipes, Lachnoclostridium, Ruthenibacterium, Lactobacillus*, and *Blautia* were frequently crucial for predictive for disease model at genus level in KRAKEN2 results with at least 92% accuracy. Moreover, these sequences were taken from V3-V5 regions, so we perform random forest for both species and genus level for BLAST-N results. *Blautia, Pedobacter, Bacteroides, Tidjanibacter, Kineothrix, Robinsoniella, Flavobacterium, Anaerostipes*, and *Gordonibacter* were frequently important for a predictive model at the genus level. The testing data accuracy was 75%. Furthermore, *Blautia wexlerae, Eubacterium saphenum, Pseudomonas migulae, Pseudomonas helleri*, and *Blautia luti* were frequently found significant for random forest results at the species level with at least 91% accuracy.

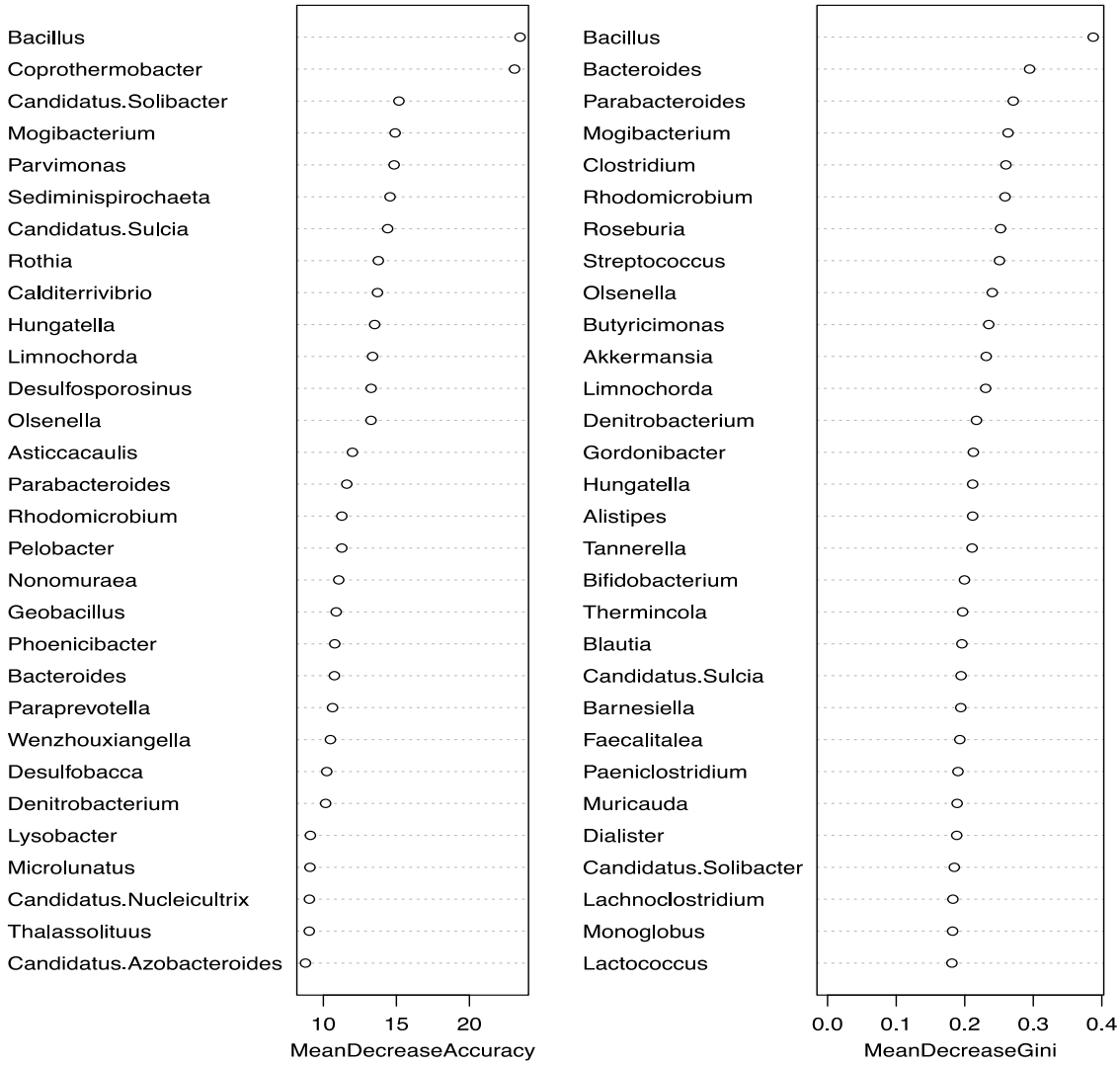

Figure 4.9. *Random Forest for first data in species-level at 60% accuracy with KRAKEN2 result.*

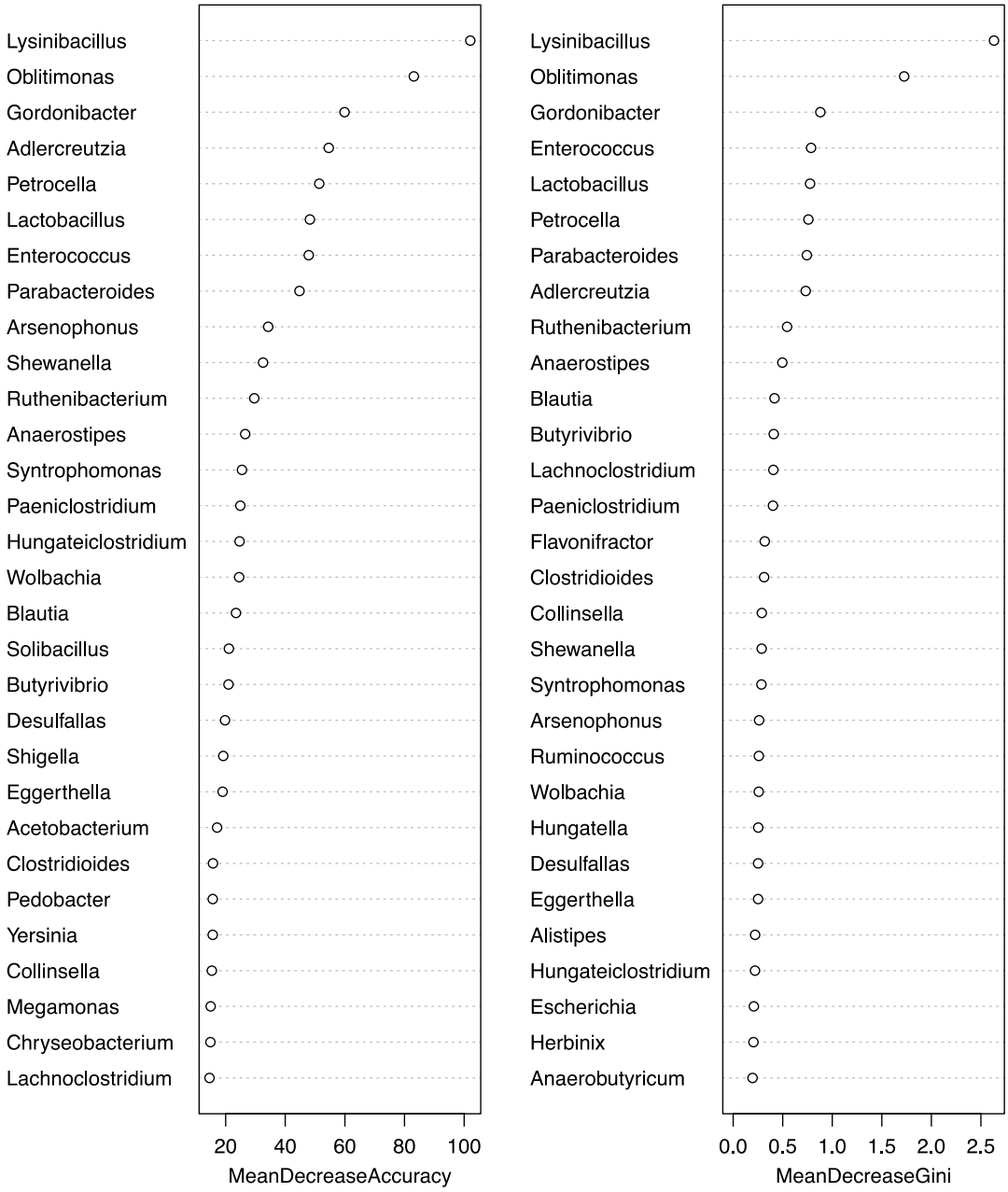

Figure 4.10. *Random Forest for second data in genus-level at 92% accuracy with KRAKEN2 result.*

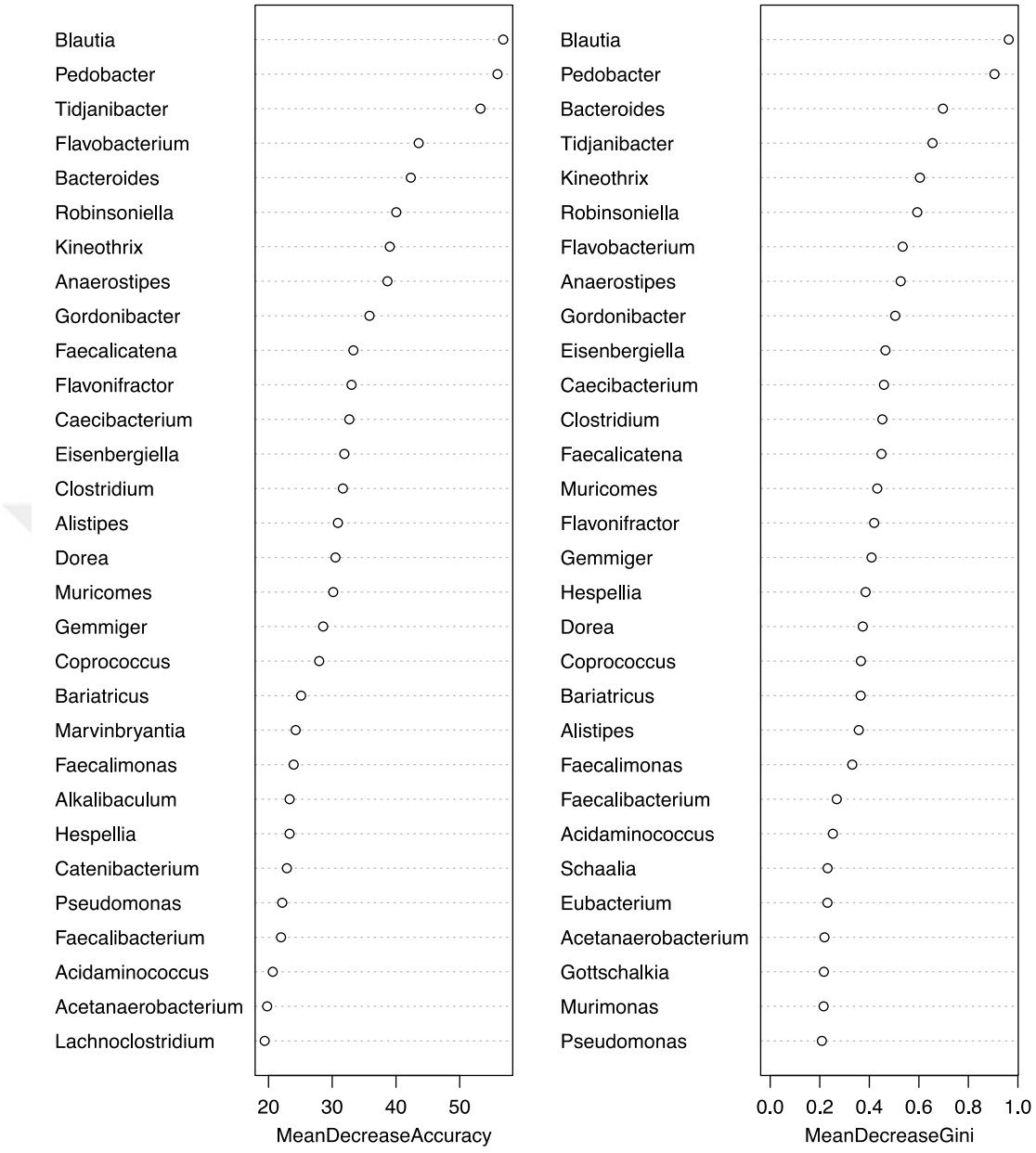

Figure 4.11. *Random Forest for second data in genus-level at 75% accuracy with BLASTn result.*

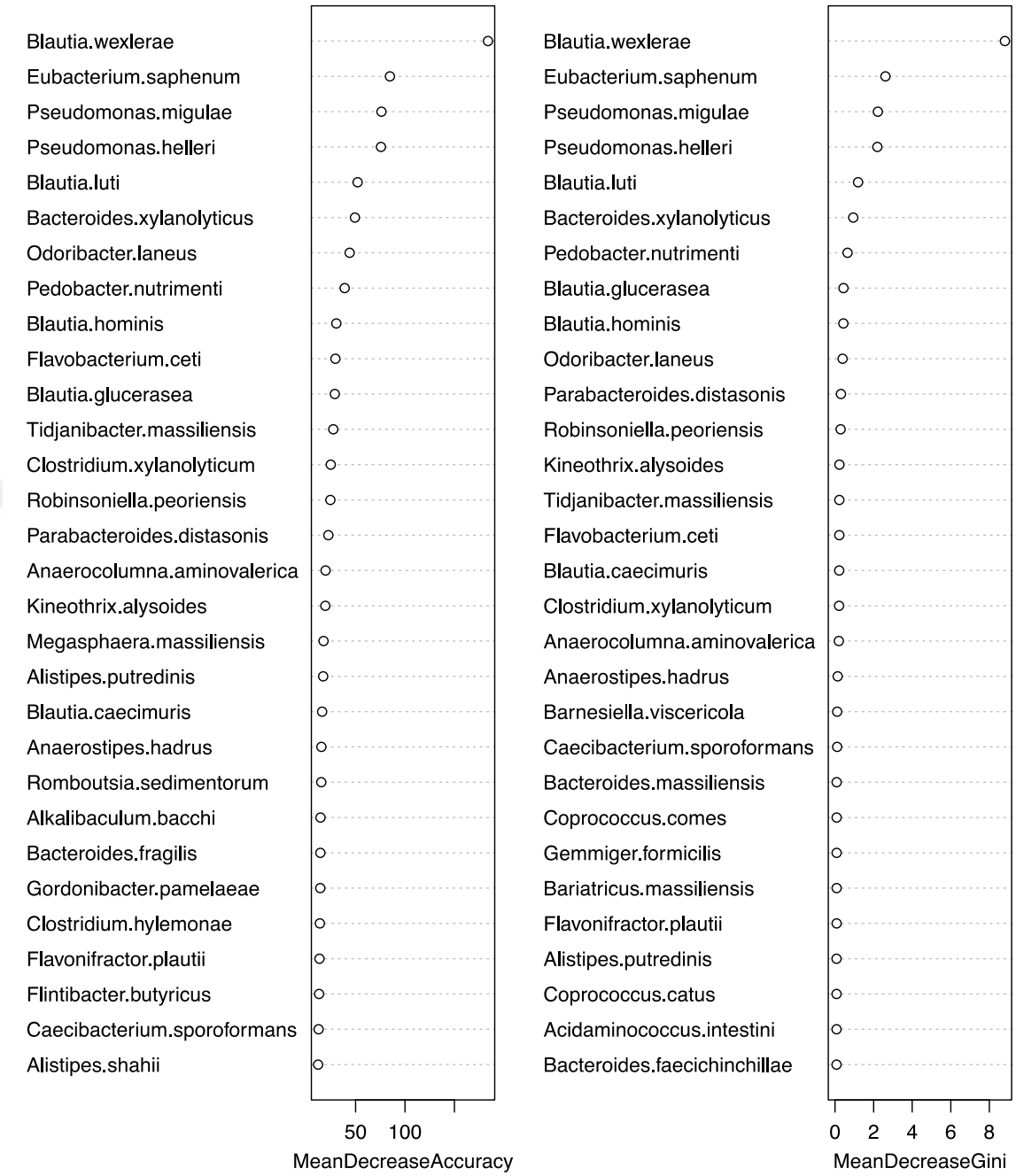

Figure 4.12. *Random Forest for second data in species-level at 92% accuracy with BLAST-N result.*

4.5 Functional Analysis

For functional analysis, we have analyzed the Human Microbiome Project datasets, which consist of bacterial genomes of the human microbiome. Then, GLIMMER was used to find protein from these bacteria datasets, and it was filtered according to our first study of bacteria data. After filtration, bacterial protein data were normalized based on bacterial abundance. We used GhostKoala to create KEGG Orthology (KO) annotations to track the differential abundance of specific functional categories across to our datasets. We compared 291 KEGG pathways and identified 39 differentially abundant KEGG pathways that show a diverse change in RRMS microbiota contrast to healthy controls. The most KEGG pathways are Amyotrophic lateral sclerosis, lipopolysaccharide biosynthesis (LPS), glutathione metabolism, lipoarabinomannan, and mineral absorption.

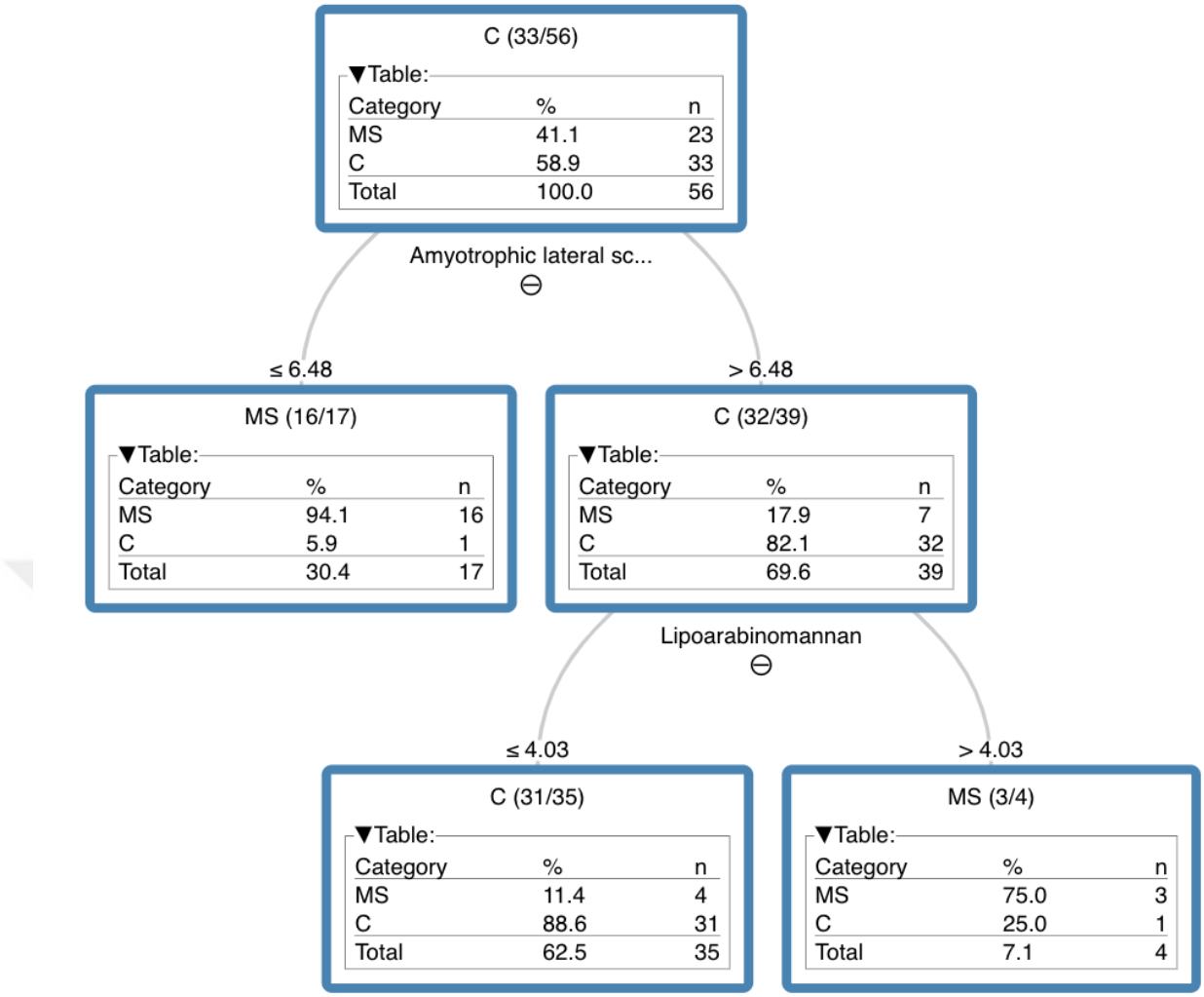

Figure 4.13. *Decision Tree for first data in genus-level at 85% accuracy with GhostKoala result. C: Control, MS: Multiple Sclerosis percentages in overall concentration.*

5. DISCUSSION and CONCLUSION

The metagenomic analysis allows scientists a deep understanding of the uncultured microbial community. The human gut microbiota is one of the most studied microbial communities in order to show interactions between host-commensal and their roles in many diseases, including autoimmune disorder, metabolic diseases, and neurodegenerative disease. For taxonomic analysis of microbiome, 16S rRNA sequencing is the most common method. However, this method has many problems; mainly the use of different techniques, low coverage of 16S rRNA regions, 16S primer set biases towards certain species between primer manufacturers and unclassified taxonomic reference datasets. Here, on behalf of showing alterations in the gut microbiome of individuals belonging to Multiple sclerosis (MS) and healthy control group, and also several biases in laboratory studies, we use two different microbiome studies; a study using only sequences of V4 regions, and another study using sequences of V3-V5 regions of 16S rRNA.

Our study demonstrates that the diversity of the fecal microbial community of RRMS patients shows distinct contrasts when compared to healthy controls. The first dataset, which contains sequences of V4 regions and analyzed in KRAKEN2, reveals the diversity at the species level with increases in *Ruthenibacterium lactatiformans, Faecalitalea cylindroides,* and *Rhodomicrobium vannielii*, and decreases in *Bacteroides sp., Paraprevotella xylaniphila, Alkaliphilus metalliredigens*, and *Caproiciproducens sp. NJN50* in RRMS patients compared to controls. These sequences were also BLASTED with 90% identity cut-off against NCBI nucleotide archive (NT) datasets. These results show that *Akkermansia muciniphila, Eubacterium ventriosum, Clostridium celerecrescens, Granulicatella adiacens,* and *Alicyclobacillus kakegawensis* species elevated abundances in RRMS patients. The differences between the results of the two methods may be due to the use of different reference databases.

In this study, various bacteria were identified to play a role in Multiple sclerosis and associated with autoimmunity defect, inflammation, and neurodegenerative disorder. We have found a higher abundance of *Rhodomicrobium vannielii*. This species oxidizes ferrous iron in the presence of light, can photo-catabolize aromatic compounds, and grows on benzoate such as bog as sole carbon source (92). This bacterium does not live in the human body, so there should be a problem with either reference datasets used by KRAKEN2, such that it should be a different species belonging to *Hyphomicrobiaceae* family.

Another most abundant species is *Ruthenibacterium lactatiformans*. It is a member of the family *Ruminococcaceae* and known for producing lactate (93). In a study conducted by Ameroni *et al*. in 2014, lactate abundance was found 2.8 times higher in MS patients than control groups, and its abundance was associated with the disability (94). In a 2019 review, *Ruthenibacterium lactatiformans* and *Paraprevotella xylaniphila* are rarely observed in human fecal microbiota. In the same review, it is associated with increased efficacy of immune checkpoint inhibitor therapy (ICT) in mice models with MC38 tumors (95). In our study, *Paraprevotella xylaniphila*, which produces succinic and acetic acids as metabolic end product from peptone, yeast extract, and glucose broth, was detected in low amounts (96). Succinic acid regulates IL-1β expression, HIF-1α activity, and ROS production, so it has an essential effect on major inflammatory pathways (97). *Faecalitalea cylindroides*, a lactic acid producer found in mouse feces, is also one of the core species in the human gut microbiome (98). Recent studies have shown that reduction in *Bacteroides* species was observed at gut microbiota of patients with MS and IBD (99). Our BLAST-N results show that *Akkermansia muciniphila* level is higher in MS patients. Other current studies illustrated that the level of this species was increased in MS groups. It is known that this species triggers proinflammatory responses in human peripheral blood mononuclear cells and monocolonized mice (100).

We also analyzed the first dataset at the genus level and found out that *Akkermansia, Anaerostipes, Mogibacterium*, and *Streptococcus* species were increased, while *Bacillus* and *Anaerotignum* were decreased in RRMS patient groups compared to control members according to KRAKEN2 results. As mentioned above, *Akkermansia muciniphila* has a relationship with Multiple sclerosis and degrades mucin. In a study, it is showed that degradation of mucin by *Akkermansia muciniphila* primarily assists the growth of *Anaerostipes caccae* and contributes butyrate development through the acetyl-CoA pathway. Moreover, they indicated that *Anaerostipes caccae* induced *Akkermansia muciniphila* by changing transcriptional response with increased expression of genes involved in mucin degradation and decreased ribosomal gene expression (101). *Mogibacterium* is mainly involved in oral microbiota, and its abundance is proportional to endodontic infections (102). Another study showed that the *Mogibacterium* abundance level was reduced in piglets, which were treated with beneficial prebiotic supplementation (103). Moreover, we have found high levels of *Streptococcus* species in the MS group. A recent study showed that when HIV-positive children were treated with cotrimoxazole, the abundance of *Streptococcus* genus in their gut microbiome and systemic inflammatory markers were both decreased (104). Our first data analysis result shows that *Bacillus* and *Anaerotignum* abundance levels were also decreased in MS patients. Current review articles mention that *Bacillus* strains produce antimicrobial peptides and small extracellular effector molecules. These molecules may have biotherapeutic potential in the food and pharmaceutical industry (105). In a study with metastatic melanoma patients receiving immune checkpoint blockade (CICB), *Bacteroides* intestinalis and *Anaerotignum lactatifermentans* presence demonstrated high or low immune-related adverse events (106).

The second dataset constructed on V3-V5 amplicon that has higher confidence in species-level showed with an increased frequency of *Blautia wexlera* in RRMS according to BLAST-N results. Moreover, *Blautia wexlerae, Eubacterium saphenum,*

Pseudomonas migulae, *Pseudomonas helleri*, and *Blautia luti* were observed as essential features for random forest results at the species level. Most of the studies exhibited increased *Blautia* and *Pseudomonas* abundances in Multiple sclerosis patients (107). Moreover, *Lysinibacillus* abundance was higher in RRMS patients compared to control groups in the secondary data. One study pointed out that *Lysinibacillus* caused bacteremia in an immunocompetent patient with no implanted prosthesis or intravascular catheters (108). And it is known that active or relapsing-remitting Multiple sclerosis lesion has tight junction abnormalities, which may be associated with inflammation and contribute to bacteremia (109). Another finding of our study is that the abundance of *Butyrivibrio*, which produces naturally conjugated linoleic acid (CLA) through the fermentation of unsaturated fatty acids in the lumen, was decreased in MS patients (110). In a study of CLA supplemented fifteen MS patients, CLA supplementation reduced inflammation in Central Nervous System (CNS) as well as increased T cell frequencies and affected activity of a myeloid cell (111). Moreover, *Dethiosulfovibrio* abundance was shown high in MS groups, whereas another study showed increased levels of this genus in rats with vitamin D deficiency (112). A recent study indicated that higher vitamin D levels reduced Multiple sclerosis symptoms (112). Likewise, *Enterobacter* levels were increased in the patient group in the second data results. *Enterobacteriaceae* showed higher endotoxin levels due to its lipopolysaccharides (LPS), which is inducing and worsening the inflammation (113).

In our study, *Parabacteroides* was determined as a significant feature to distinguish MS patients. Recent studies show that *Parabacteroides* relieve obesity and metabolic disorder by producing succinate and bile acid in the mouse gut microbiome (114). Moreover, Parabacteroides have shown lower abundance in the gut microbiome of Multiple sclerosis patients. This genus promotes anti-inflammatory IL-10–expressing human CD4+CD25+ T cells and IL-10+FoxP3+ regulatory T cell (Tregs) in mice (100). In contrast, another study demonstrated that *Parabacteroides* species produced antagonistic substances in the human gut, and this production may be used against microorganisms in the gut, such as gram-positive bacteria (115). Recent studies

displayed that *Prevotella, Parabacteroides, Adlercreutzia, Slackia,* and *Lactobacillus,* whose members can metabolize phytoestrogen compounds into beneficial metabolites were found in reduced levels in MS patients (116). *Adlercreutzia equolifaciens* and *Slackia* produce equol, a nonsteroidal estrogen in the human gut (117,118). Furthermore, the main milk fermenter *Lactobacillus* can activate anti-inflammatory effects on immune cells by bioconversion of phytoestrogen compounds (119). There is a study about Quercetin, which is a flavonoid phytoestrogen, and its treatment enhances mice with Experimental allergic encephalomyelitis (EAE) by inhibiting IL-12 production and activating T cells (120).

Our findings also show that these bacteria are distinguishing features for Multiple sclerosis data, which also supports the claim that phytoestrogens and their metabolites may induce its anti-inflammatory effects in patients with autoimmune diseases such as Multiple sclerosis. Moreover, we also determined the most critical KEGG pathways, including Amyotrophic lateral sclerosis, lipopolysaccharide biosynthesis, glutathione metabolism, lipoarabinomannan, and mineral absorption. In a previous study, homeostasis of Glutathione (GSH), the major antioxidant in the brain, is changed in MS patients (121). Werner et al. showed that glutaminase, which converts glutamine to glutamate, was expressed in high amounts. This production may govern the degeneration of axons and death of oligodendrocyte in mouse with MS (122). And also, the lipid A structure of LPS was associated with immune recognition of signaling through and the Toll-like receptor 4 (TLR4) complex activation (123). Our approach is empirical, not experimental, because it based on the abundances of bacteria instead of protein expression levels. This analysis should also be further supported by meta-transcriptome analysis.

In this thesis, there are some limitations including primer biases and taxonomic coverage of different 16S subregions, geographical location, incorrect or unclassified

taxonomic reference sequences, run costs of whole-genome sequencing approach, and sequencing techniques may because of different results in many microbiome studies. One of the main disadvantages is size of 16S rRNA regions. 16S rRNA regions are approximately 1600 base pairs long, and they are the most significant factor for the identification of community composition. Although the sequencing cost of extended read length is pricy, it can span multiple hypervariable regions of the 16S rRNA gene, which then can be taxonomically classified at a species level. The short-read sequencing platforms have a large amount of PCR primers to amplify 16S rRNA regions of different hypervariable regions for sequencing (124). The microbiome sampling location is also affecting library preparation. In order to observe better taxonomic profiles, researchers might require different regions of 16S amplicons for library preparations concerning the sampling regions of the body. Researchers of Yu. S. Bukin *et al.* compared V2-V3 and V3-V4 amplicons, where the V2-V3 regions have higher resolution at the genus and species level than the V3-V4 region. Another problem is DNA contamination from DNA extraction kits and other laboratory reagents affects PCR-based amplicons and shotgun metagenomics sequencing results (125). Other important problem for metagenomic analysis is 16S rRNA gene database constraints. These databases contain 1.8 million bacteria and archaea at different taxonomic levels, and they also include incomplete information for some sequences, especially at lower taxonomic levels such as species. The bioinformatics tools generally use operational taxonomic units (OTUs) to cluster sequencing reads K00 for the genus level (126). The most popular and faster taxonomic annotation programs are based on k-mer, map, and BLAST. They all have some weaknesses and strengths while constructing OTU tables. Here is a handful of examples of observed taxonomic analysis problems; both the complete 16S sequence identity and shorter fragments of 16S regions might not be sufficient enough to have higher confidence. For example, the complete 16S sequence identity between *E. coli* (NR_024570.1) and *Shigella flexneri* (NR_026331.1) is 99,24%. The observed identity reaches 99,76 % whenever the amplicon region is V3-V4. Through the evolution, microorganisms have entropic adaptations where they have gained multiple copies of transcriptional and translational machinery elements to have improved adaptation for environmental changes. The current bacterial database shows us that more than 86% of the bacteria have multiple

copies of 16S regions. Not only are amplicons showing biased distributions, but also multiple copies bearing few SNPs on them make indistinguishable OTUs. While *Shigella Flexneri* (MN960402.1), *Citrobacter freundii* (MH169223), *Shigella sonnei* (CP045524.1) and *Salmonella sp S13* (CP047094.1) are sharing 99.76 percent similarity, the *Escherichia coli* (NR_114042.1) and *Escherichia marmotae* (NR_136472.1) are sharing 98.82 sequence similarity as in the case of the V3-V4 region. Observing higher identities in different genus levels and lesser identities in species levels might neutralize the k-mer based taxonomy programs that can be of great help. Although the mean Illumina read quality scores are in between q28-q34 for microbiome libraries, human gut microbiome studies as V3-V4 region can complicate researchers with PCR bias derived from primer sensitivity even with an unbiased DNA isolation.

MS researchers will be investigating disease management with enhanced therapeutics with ample help from microbiome modulation, using current analysis. While our findings require further analysis of the wet lab, our current results will be used for adjunctive therapies related to microbiome modulation and new project proposals will be ignited.

6. REFERENCES

1. Polman CH, Reingold SC, Banwell B, Clanet M, Cohen JA, Filippi M, et al. Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. Ann Neurol. 2011 Feb;69(2):292–302.

2. Reipert B. Multiple sclerosis: a short review of the disease and its differences between men and women. J Mens Health Gend. 2004 Dec 1;1(4):334–40.

3. Organization WH, Federation MSI. Atlas : multiple sclerosis resources in the world 2008. Geneva PP - Geneva: World Health Organization;

4. Goldenberg MM. Multiple sclerosis review. P T Peer-Rev J Formul Manag. 2012 Mar;37(3):175–84.

5. Olsson T, Barcellos LF, Alfredsson L. Interactions between genetic, lifestyle and environmental risk factors for multiple sclerosis. Nat Rev Neurol. 2017 Jan;13(1):25–36.

6. Chen J, Chia N, Kalari KR, Yao JZ, Novotna M, Paz Soldan MM, et al. Multiple sclerosis patients have a distinct gut microbiota compared to healthy controls. Sci Rep. 2016;6(1):28484.

7. Mallick H, Franzosa EA, Mclver LJ, Banerjee S, Sirota-Madi A, Kostic AD, et al. Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences. Nat Commun. 2019;10(1):3136.

8. Stangel M, Kleinschnitz C, Mäurer M, Raab P, Sühs K-W, Trebst C. Multiple Sklerose und andere autoimmune ZNS-Erkrankungen BT - Autoimmunerkrankungen in der Neurologie: Diagnostik und Therapie. In: Stangel M, Mäurer M, editors. Berlin, Heidelberg: Springer Berlin Heidelberg; 2018. p. 1–103.

9. Hafler DA, Compston A, Sawcer S, Lander ES, Daly MJ, De Jager PL, et al. Risk alleles for multiple sclerosis identified by a genomewide study. N Engl J Med. 2007 Aug;357(9):851–62.

10. Zhou Y, Graves JS, Simpson SJ, Charlesworth JC, Mei I van der, Waubant E, et al. Genetic variation in the gene LRP2 increases relapse risk in multiple sclerosis. J Neurol Neurosurg Psychiatry. 2017 Oct;88(10):864–8.

11. Hilven K, Vandebergh M, Smets I, Mallants K, Goris A, Dubois B. Genetic basis for relapse rate in multiple sclerosis: Association with LRP2 genetic variation. Mult Scler Houndmills Basingstoke Engl. 2018 Nov;24(13):1773–5.

12. Graves JS, Barcellos LF, Simpson S, Belman A, Lin R, Taylor B V, et al. The multiple sclerosis risk allele within the AHI1 gene is associated with relapses in children and adults. Mult Scler Relat Disord. 2018 Jan;19:161–5.

13. Aslani S, Jafari N, Javan MR, Karami J, Ahmadi M, Jafarnejad M. Epigenetic Modifications and Therapy in Multiple Sclerosis. Neuromolecular Med. 2017 Mar;19(1):11–23.

14. Graves MC, Benton M, Lea RA, Boyle M, Tajouri L, Macartney-Coxson D, et al. Methylation differences at the HLA-DRB1 locus in CD4+ T-Cells are associated with multiple sclerosis. Mult Scler Houndmills Basingstoke Engl. 2014 Jul;20(8):1033–41.

15. Pedre X, Mastronardi F, Bruck W, López-Rodas G, Kuhlmann T, Casaccia P. Changed histone acetylation patterns in normal-appearing white matter and early multiple sclerosis lesions. J Neurosci Off J Soc Neurosci. 2011 Mar;31(9):3435–45.

16. Aung LL, Mouradian MM, Dhib-Jalbut S, Balashov KE. MMP-9 expression is increased in B lymphocytes during multiple sclerosis exacerbation and is regulated by microRNA-320a. J Neuroimmunol. 2015;278:185—189.

17. Olsson T, Barcellos LF, Alfredsson L. Interactions between genetic, lifestyle and environmental risk factors for multiple sclerosis. Nat Rev Neurol. 2017 Jan;13(1):25–36.

18. Salzer J, Hallmans G, Nyström M, Stenlund H, Wadell G, Sundström P. Smoking as a risk factor for multiple sclerosis. Mult Scler Houndmills Basingstoke Engl. 2013 Jul;19(8):1022–7.

19. Zhou Y, Zhu G, Charlesworth JC, Simpson SJ, Rubicz R, Göring HH, et al. Genetic loci for Epstein-Barr virus nuclear antigen-1 are associated with risk of multiple sclerosis. Mult Scler Houndmills Basingstoke Engl. 2016 Nov;22(13):1655–64.

20. Farrokhi V, Nemati R, Nichols FC, Yao X, Anstadt E, Fujiwara M, et al. Bacterial lipodipeptide, Lipid 654, is a microbiome-associated biomarker for multiple sclerosis. Clin Transl Immunol. 2013 Nov;2(11):e8.

21. Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, et al. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486(7402):207–14.

22. Methé BA, Nelson KE, Pop M, Creasy HH, Giglio MG, Huttenhower C, et al. A framework for human microbiome research. Nature. 2012;486(7402):215–21.

23. Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch S V, Knight R. Current understanding of the human microbiome. Nat Med. 2018;24(4):392–400.

24. Ni J, Shen T-CD, Chen EZ, Bittinger K, Bailey A, Roggiani M, et al. A role for bacterial urease in gut dysbiosis and Crohn's disease. Sci Transl Med. 2017 Nov 15;9(416):eaah6888.

25. Kostic AD, Chun E, Robertson L, Glickman JN, Gallini CA, Michaud M, et al. Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. Cell Host Microbe. 2013 Aug 14;14(2):207–15.

26. Petersen C, Round JL. Defining dysbiosis and its influence on host immunity and disease. Cell Microbiol. 2014 Jul;16(7):1024–33.

27. Forsythe P, Kunze W, Bienenstock J. Moody microbes or fecal phrenology: what do we know about the microbiota-gut-brain axis? BMC Med. 2016;14(1):58.

28. Sarkar A, Lehto SM, Harty S, Dinan TG, Cryan JF, Burnet PWJ. Psychobiotics and the Manipulation of Bacteria-Gut-Brain Signals. Trends Neurosci. 2016/10/25 ed. 2016 Nov;39(11):763–81.

29. Young SN. Acute tryptophan depletion in humans: a review of theoretical, practical and ethical aspects. J Psychiatry Neurosci JPN. 2013 Sep;38(5):294–305.

30. Julio-Pieper M, O'Connor RM, Dinan TG, Cryan JF. Regulation of the brain-gut axis by group III metabotropic glutamate receptors. Eur J Pharmacol. 2013 Jan;698(1–3):19–30.

31. Sano C. History of glutamate production. Am J Clin Nutr. 2009 Jul 29;90(3):728S-732S.

32. Tanous C, Chambellon E, Sepulchre A-M, Yvon M. The gene encoding the glutamate dehydrogenase in Lactococcus lactis is part of a remnant Tn3 transposon carried by a large plasmid. J Bacteriol. 2005;187(14):5019—5022.

33. Mazzoli R, Pessione E. The Neuro-endocrinological Role of Microbial Glutamate and GABA Signaling . Vol. 7, Frontiers in Microbiology . 2016. p. 1934.

34. Barrett E, Ross RP, O'Toole PW, Fitzgerald GF, Stanton C. γ-Aminobutyric acid production by culturable bacteria from the human intestine. J Appl Microbiol. 2012 Aug 1;113(2):411–7.

35. Siragusa S, De Angelis M, Di Cagno R, Rizzello CG, Coda R, Gobbetti M. Synthesis of gamma-aminobutyric acid by lactic acid bacteria isolated from a variety of Italian cheeses. Appl Environ Microbiol. 2007/09/21 ed. 2007 Nov;73(22):7283–90.

36. Chang P V, Hao L, Offermanns S, Medzhitov R. The microbial metabolite butyrate regulates intestinal macrophage function via histone deacetylase inhibition. Proc Natl Acad Sci U S A. 2014 Feb;111(6):2247–52.

37. Sharon G, Garg N, Debelius J, Knight R, Dorrestein PC, Mazmanian SK. Specialized Metabolites from the Microbiome in Health and Disease. Cell Metab. 2014;20(5):719–30.

38. Armougom F, Henry M, Vialettes B, Raccah D, Raoult D. Monitoring Bacterial Community of Human Gut Microbiota Reveals an Increase in Lactobacillus in Obese Patients and Methanogens in Anorexic Patients. PLOS ONE. 2009 Sep 23;4(9):e7125.

39. Round JL, Mazmanian SK. The gut microbiota shapes intestinal immune responses during health and disease. Nat Rev Immunol. 2009 May;9(5):313–23.

40. Wang Y, Kasper LH. The role of microbiome in central nervous system disorders. Brain Behav Immun. 2014 May;38:1–12.

41. Wang B, Jiang X, Cao M, Ge J, Bao Q, Tang L, et al. Altered Fecal Microbiota Correlates with Liver Biochemistry in Nonobese Patients with Non-alcoholic Fatty Liver Disease. Sci Rep. 2016 Aug;6:32002.

42. Pickard JM, Zeng MY, Caruso R, Núñez G. Gut microbiota: Role in pathogen colonization, immune responses, and inflammatory disease. Immunol Rev. 2017 Sep;279(1):70–89.

43. Erny D, Hrabě de Angelis AL, Jaitin D, Wieghofer P, Staszewski O, David E, et al. Host microbiota constantly control maturation and function of microglia in the CNS. Nat Neurosci. 2015 Jul;18(7):965–77.

44. Westfall S, Lomis N, Kahouli I, Dia S, Singh S, Prakash S. Microbiome, probiotics and neurodegenerative diseases: deciphering the gut brain axis. Cell Mol Life Sci CMLS. 2017 Jun 22;74.

45. 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.

46. Macfarlane GT, Macfarlane S. Human Colonic Microbiota: Ecology, Physiology and Metabolic Potential of Intestinal Bacteria. Scand J Gastroenterol. 1997 Jan 1;32(sup222):3–9.

47. Rivière A, Selak M, Lantin D, Leroy F, De Vuyst L. Bifidobacteria and Butyrate-Producing Colon Bacteria: Importance and Strategies for Their Stimulation in the Human Gut. Front Microbiol. 2016 Jun 28;7:979.

48. Zhang M, Qiu X, Zhang H, Yang X, Hong N, Yang Y, et al. Faecalibacterium prausnitzii inhibits interleukin-17 to ameliorate colorectal colitis in rats. PloS One. 2014 Oct 2;9(10):e109146–e109146.

49. Scheiman J, Luber J, Chavkin T, Macdonald T, Tung A, Wibowo M, et al. Meta-omics analysis of elite athletes identifies a performance-enhancing microbe that functions via lactate metabolism. Nat Med. 2019 Jul 1;25.

50. Lundberg JO, Weitzberg E, Cole JA, Benjamin N. Nitrate, bacteria and human health. Nat Rev Microbiol. 2004 Jul;2(7):593–602.

51. Duncan C, Dougall H, Johnston P, Green S, Brogan R, Leifert C, et al. Chemical generation of nitric oxide in the mouth from the enterosalivary circulation of dietary nitrate. Nat Med. 1995 Jun;1(6):546–51.

52. Hyde ER, Andrade F, Vaksman Z, Parthasarathy K, Jiang H, Parthasarathy DK, et al. Metagenomic analysis of nitrate-reducing bacteria in the oral cavity: implications for nitric oxide homeostasis. PloS One. 2014;9(3):e88645.

53. Carlström M, Persson AEG, Larsson E, Hezel M, Scheffer PG, Teerlink T, et al. Dietary nitrate attenuates oxidative stress, prevents cardiac and renal injuries, and reduces blood pressure in salt-induced hypertension. Cardiovasc Res. 2011 Feb;89(3):574–85.

54. Bryan NS, Calvert JW, Elrod JW, Gundewar S, Ji SY, Lefer DJ. Dietary nitrite supplementation protects against myocardial ischemia-reperfusion injury. Proc Natl Acad Sci U S A. 2007 Nov;104(48):19144–9.

55. Parsonnet J, Friedman GD, Vandersteen DP, Chang Y, Vogelman JH, Orentreich N, et al. Helicobacter pylori infection and the risk of gastric carcinoma. N Engl J Med. 1991 Oct;325(16):1127–31.

56. Karlsson FH, Tremaroli V, Nookaew I, Bergström G, Behre CJ, Fagerberg B, et al. Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature. 2013;498(7452):99–103.

57. Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, et al. Article: Richness of human gut microbiome correlates with metabolic markers E. Le Chatelier, T. Nielsen, J. Qin, E. Prifti, F. Hildebrand, G. Falony, M. Almeida, M. Arumugam, J. M. Batto, S. Kennedy, [.....], M. Derrien, J. E. van Hylckama Vlieg, P. Veig. 2013 Jan 1;

58. Fearon RMP, Reiss D, Leve LD, Shaw DS, Scaramella L V., Ganiban JM, et al. Host genotype and gut microbiome modulate insulin secretion and diet-induced metabolic phenotypes. Cell Rep. 2015;27(4):1251–65.

59. Handelsman Y. Role of bile acid sequestrants in the treatment of type 2 diabetes. Diabetes Care. 2011 May;34 Suppl 2(Suppl 2):S244–50.

60. Heiss CN, Olofsson LE. Gut Microbiota-Dependent Modulation of Energy Metabolism. J Innate Immun. 2018;10(3):163–71.

61. Gomez-Arango LF, Barrett HL, McIntyre HD, Callaway LK, Morrison M, Dekker Nitert M. Connections Between the Gut Microbiome and Metabolic

Hormones in Early Pregnancy in Overweight and Obese Women. Diabetes. 2016 Aug 1;65(8):2214 LP – 2223.

62. Gonzalez A, Navas-Molina JA, Kosciolek T, McDonald D, Vázquez-Baeza Y, Ackermann G, et al. Qiita: rapid, web-enabled microbiome meta-analysis. Nat Methods. 2018 Oct;15(10):796–8.

63. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, et al. A communal catalogue reveals Earth's multiscale microbial diversity. Nature. 2017;551(7681):457–63.

64. Liu EW, Jansson JK, Genomics C, History M, Nelson PKE, Bryan PA, et al. Environmental molecular microbiology. Int Microbiol - Off J Span Soc Microbiol. 2009;12(4):254–5.

65. Janda JM, Abbott SL. 16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls. J Clin Microbiol. 2007 Sep 1;45(9):2761 LP – 2764.

66. Bikel S, Valdez-Lara A, Cornejo-Granados F, Rico K, Canizales-Quinteros S, Soberón X, et al. Combining metagenomics, metatranscriptomics and viromics to explore novel microbial interactions: towards a systems-level understanding of human microbiome. Comput Struct Biotechnol J. 2015;13:390–401.

67. Cox MJ, Cookson WOCM, Moffatt MF. Sequencing the human microbiome in health and disease. Hum Mol Genet. 2013 Aug 13;22(R1):R88–94.

68. Stackebrandt E, Goebel BM. Taxonomic note: A place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. Int J Syst Bacteriol. 1994;44(4):846–9.

69. Ashelford KE, Chuzhanova NA, Fry JC, Jones AJ, Weightman AJ. At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. Appl Environ Microbiol. 2005 Dec;71(12):7724–36.

70. Pallen MJ, Wren BW. Bacterial pathogenomics. Nature. 2007 Oct;449(7164):835–42.

71. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol. 2007 Aug;73(16):5261–7.

72. Vinje H, Almøy T, Liland KH, Snipen L. A systematic search for discriminating sites in the 16S ribosomal RNA gene. Microb Inform Exp. 2014 Jan;4(1):2.

73. Ounit R, Wanamaker S, Close TJ, Lonardi S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics. 2015;16(1):236.

74. McHardy AC, Martín HG, Tsirigos A, Hugenholtz P, Rigoutsos I. Accurate phylogenetic classification of variable-length DNA fragments. Nat Methods. 2007;4(1):63–72.

75. Chatterji S, Yamazaki I, Bai Z, Eisen JA. CompostBin: A DNA Composition-Based Algorithm for Binning Environmental Shotgun Reads BT - Research in Computational Molecular Biology. In: Vingron M, Wong L, editors. Berlin, Heidelberg: Springer Berlin Heidelberg; 2008. p. 17–28.

76. Diaz NN, Krause L, Goesmann A, Niehaus K, Nattkemper TW. TACOA: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. BMC Bioinformatics. 2009 Feb;10:56.

77. Zheng H, Wu H. Short Prokaryotic DNA Fragment Binning Using a Hierarchical Classifier Based on Linear Discriminant Analysis and Principal Component Analysis. J Bioinform Comput Biol. 2010 Dec 1;8:995–1011.

78. Wood D, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019 Dec 1;20.

79. Noé L, Noé L, Martin DEK. A coverage criterion for spaced seeds and its applications to support vector machine string kernels and k-mer distances. J Comput Biol J Comput Mol Cell Biol. 2014;21(12):947—963.

80. Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A. 1977;74(11):5088—5090.

81. Xia LC, Cram JA, Chen T, Fuhrman JA, Sun F. Accurate genome relative abundance estimation based on shotgun metagenomic reads. PloS One. 2011;6(12):e27992.

82. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Altschul_1990_5424.pdf. Vol. 215, Journal of Molecular Biology. 1990. p. 403–10.

83. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015 Jun;16(6):321–32.

84. Gareth James Trevor Hastie, Robert Tibshirani DW. An introduction to statistical learning : with applications in R. New York : Springer, [2013] ©2013;

85. Ripley BD. Pattern recognition and neural networks. Pattern Recognit Neural Netw. 2014;(1995):1–403.

86. Brownlee J. What is the difference between test and validation datasets. Mach Learn Mastery Httpsmachinelearningmastery Comdifference-Test-Valid-Datasets Accessed Nov. 2017;

87. Zhao Y, Zhang Y. Comparison of decision tree methods for finding active objects. Adv Space Res. 2008;41(12):1955–9.

88. Breiman L. Machine Learning, Volume 45, Number 1 - SpringerLink. Mach Learn. 2001 Oct 1;45:5–32.

89. Breiman L. Bagging Predictors. Mach Learn. 1996;24(2):123–40.

90. Qi Y. Random forest for bioinformatics. Ensemble Mach Learn Methods Appl. 2012;307–23.

91. Pasolli E, Truong DT, Malik F, Waldron L, Segata N. Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. PLOS Comput Biol. 2016 Jul 11;12(7):e1004977.

92. Wright G, Madigan M. Photocatabolism of Aromatic Compounds by the Phototrophic Purple Bacterium Rhodomicrobium vannielii. Appl Environ Microbiol. 1991 Aug 1;57:2069–73.

93. Shkoporov AN, Chaplin A V., Shcherbakova VA, Suzina NE, Kafarskaia LI, Bozhenko VK, et al. Ruthenibacterium lactatiformans gen. nov., sp. nov., an anaerobic, lactate-producing member of the family Ruminococcaceae isolated from human faeces. Int J Syst Evol Microbiol. 2016;66(8):3041–9.

94. Amorini AM, Nociti V, Petzold A, Gasperini C, Quartuccio E, Lazzarino G, et al. Serum lactate as a novel potential biomarker in multiple sclerosis. Biochim Biophys Acta BBA - Mol Basis Dis. 2014;1842(7):1137–43.

95. Frankel AE, Deshmukh S, Reddy A, Lightcap J, Hayes M, McClellan S, et al. Cancer Immune Checkpoint Inhibitor Therapy and the Gut Microbiota. Integr Cancer Ther. 2019 Jan 1;18:1534735419846379.

96. Morotomi M, Nagai F, Sakon H, Tanaka R. Paraprevotella clara gen. nov., sp. nov. and Paraprevotella xylaniphila sp. nov., members of the family "Prevotellaceae" isolated from human faeces. Int J Syst Evol Microbiol. 2009 Aug;59(Pt 8):1895–900.

97. Kelly B, O'Neill LAJ. Metabolic reprogramming in macrophages and dendritic cells in innate immunity. Cell Res. 2015;25(7):771–84.

98. Chang D-H, Rhee M-S, Ahn S, Bang B-H, Oh JE, Lee HK, et al. Faecalibaculum rodentium gen. nov., sp. nov., isolated from the faeces of a laboratory mouse. Antonie Van Leeuwenhoek. 2015 Dec;108(6):1309–18.

99. Miyake S, Kim S, Suda W, Oshima K, Nakamura M, Matsuoka T, et al. Dysbiosis in the Gut Microbiota of Patients with Multiple Sclerosis, with a Striking Depletion of Species Belonging to Clostridia XIVa and IV Clusters. PloS One. 2015;10(9):e0137429.

100. Cekanaviciute E, Yoo BB, Runia TF, Debelius JW, Singh S, Nelson CA, et al. Gut bacteria from multiple sclerosis patients modulate human T cells and exacerbate symptoms in mouse models. Proc Natl Acad Sci. 2017 Oct 3;114(40):10713 LP – 10718.

101. Chia LW, Hornung BVH, Aalvink S, Schaap PJ, de Vos WM, Knol J, et al. Deciphering the trophic interaction between Akkermansia muciniphila and the butyrogenic gut commensal Anaerostipes caccae using a metatranscriptomic approach. Antonie Van Leeuwenhoek. 2018 Jun;111(6):859–73.

102. Saito D, De Toledo Leonardo R, Mazza Rodrigues JL, Tsai SM, Höfling JF, Gonçalves RB. Identification of bacteria in endodontic infections by sequence analysis of 16S rDNA clone libraries. J Med Microbiol. 2006;55(1):101–7.

103. O'Hara E, Kelly A, McCabe MS, Kenny DA, Guan LL, Waters SM. Effect of a butyrate-fortified milk replacer on gastrointestinal microbiota and products of fermentation in artificially reared dairy calves at weaning. Sci Rep. 2018;8(1):14901.

104. Bourke CD, Gough EK, Pimundu G, Shonhai A, Berejena C, Terry L, et al. Cotrimoxazole reduces systemic inflammation in HIV infection by altering the gut microbiome and immune activation. Sci Transl Med. 2019 Apr 3;11(486):eaav0537.

105. Elshaghabee FMF, Rokana N, Gulhane RD, Sharma C, Panwar H. Bacillus as potential probiotics: Status, concerns, and future perspectives. Front Microbiol. 2017;8(AUG):1–15.

106. Gopalakrishnan V, Andrews M, Chen W-S, Spencer C, Vence L, Reuben A, et al. Abstract 1493: Therapeutic efficacy and tolerability of combined immune checkpoint blockade in metastatic melanoma patients is influenced by the gut microbiome. Cancer Res. 2019 Jul 1;79(13 Supplement):1493 LP – 1493.

107. Shahi SK, Freedman SN, Mangalam AK. Gut microbiome in multiple sclerosis: The players involved and the roles they play. Gut Microbes. 2017 Nov;8(6):607–15.

108. Wenzler E, Kamboj K, Balada-Llasat JM. Severe Sepsis Secondary to Persistent Lysinibacillus sphaericus, Lysinibacillus fusiformis and Paenibacillus amylolyticus Bacteremia. Int J Infect Dis. 2015;35:e93–5.

109. Plumb J, McQuaid S, Mirakhur M, Kirk J. Abnormal Endothelial Tight Junctions in Active Lesions and Normal-appearing White Matter in Multiple Sclerosis. Brain Pathol. 2002 Apr 1;12(2):154–69.

110. Witkamp RF. 3.15 - Biologically Active Compounds in Food Products and Their Effects on Obesity and Diabetes. In: Liu H-W (Ben), Mander LBT-CNPII, editors. Oxford: Elsevier; 2010. p. 509–45.

111. A.-K. F, S. H, M. H, F. T, M. H, K. B, et al. Dietary supplementation with conjugated linoleic acid influences central nervous system autoimmunity via the gut-central nervous system axis. Mult Scler J. 2018;24(2 Supplement):600–1.

112. Robles-Vera I, Callejo M, Ramos R, Duarte J, Perez-Vizcaino F. Impact of vitamin D deficit on the rat gut microbiome. Nutrients. 2019;11(11).

113. Keskitalo A, Munukka E, Toivonen R, Hollmén M, Kainulainen H, Huovinen P, et al. Enterobacter cloacae administration induces hepatic damage and subcutaneous fat accumulation in high-fat diet fed mice. PloS One. 2018 May 30;13(5):e0198262–e0198262.

114. Wang K, Liao M, Zhou N, Bao L, Ma K, Zheng Z, et al. Parabacteroides distasonis Alleviates Obesity and Metabolic Dysfunctions via Production of Succinate and Secondary Bile Acids. Cell Rep. 2019;26(1):222-235.e5.

115. Nakano V, Ignacio A, R. Fernandes M, Fugukaiti M, Avila-Campos M. Intestinal Bacteroides and Parabacteroides species producing antagonistic substances. Curr Trends Microbiol. 2013 Jan 1;1–4.

116. Freedman SN, Shahi SK, Mangalam AK. The "Gut Feeling": Breaking Down the Role of Gut Microbiome in Multiple Sclerosis. Neurotherapeutics. 2018;15(1):109–25.

117. Schröder C, Matthies A, Engst W, Blaut M, Braune A. Identification and expression of genes involved in the conversion of daidzein and genistein by the equol-forming bacterium Slackia isoflavoniconvertens. Appl Environ Microbiol. 2013/03/29 ed. 2013 Jun;79(11):3494–502.

118. Maruo T, Sakamoto M, Ito C, Toda T, Benno Y. Adlercreutzia equolifaciens gen. nov., sp. nov., an equol-producing bacterium isolated from human faeces, and emended description of the genus Eggerthella. Int J Syst Evol Microbiol. 2008 May;58(Pt 5):1221–7.

119. Rekha CR, Vijayalakshmi G. Bioconversion of isoflavone glycosides to aglycones, mineral bioavailability and vitamin B complex in fermented soymilk by probiotic bacteria and yeast. J Appl Microbiol. 2010 Oct;109(4):1198–208.

120. Muthian G, Bright JJ. Quercetin, a Flavonoid Phytoestrogen, Ameliorates Experimental Allergic Encephalomyelitis by Blocking IL-12 Signaling Through JAK-STAT Pathway in T Lymphocyte. J Clin Immunol. 2004;24(5):542–52.

121. Carvalho AN, Lim JL, Nijland PG, Witte ME, Van Horssen J. Glutathione in multiple sclerosis: More than just an antioxidant? Mult Scler J. 2014 May 19;20(11):1425–31.

122. Werner P, Pitt D, Raine CS. Multiple sclerosis: altered glutamate homeostasis in lesions correlates with oligodendrocyte and axonal damage. Ann Neurol. 2001 Aug;50(2):169–80.

123. Kim HM, Park BS, Kim J-I, Kim SE, Lee J, Oh SC, et al. Crystal structure of the TLR4-MD-2 complex with bound endotoxin antagonist Eritoran. Cell. 2007 Sep;130(5):906–17.

124. Bikel S, Valdez-Lara A, Cornejo-Granados F, Rico K, Canizales-Quinteros S, Soberón X, et al. Combining metagenomics, metatranscriptomics and viromics

to explore novel microbial interactions: towards a systems-level understanding of human microbiome. Comput Struct Biotechnol J. 2015 Jan 1;13:390–401.

125. Bukin YuS, Galachyants YuP, Morozov IV, Bukin SV, Zakharenko AS, Zemskaya TI. The effect of 16S rRNA region choice on bacterial community metabarcoding results. Sci Data. 2019 ubat;6(1):190007.

126. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 2014 Nov 12;12(1):87.

CURRICULUM VITAE