Winter 1-31-2022

A Predictive Model for Geographical Location Using Microbiome Composition and COVID-19 Based Analysis

Ahmed Adel Aboushanab
ahmedadelaboushanab@aucegypt.edu

THE AMERICAN UNIVERSITY IN CAIRO
الجامعة الأمريكية بالقاهرة

School of Sciences and Engineering

A predictive model for geographical location using microbiome composition and COVID-19 based analysis

A Thesis Submitted to

Biotechnology Graduate Program

in partial fulfillment of the requirements for

the degree in Master of Science

by

Ahmed Adel Aboushanab

Bachelor of Biomedical Sciences, University of Science and Technology - Zewail City 2017

Under the supervision of:

Prof. Asma Amleh

Biology Department

The American University in Cairo

Prof. Ahmed Moustafa

Biology Department

The American University in Cairo

January /2022

Abstract

The human microbiome is a main contributor for the health and welfare of the human body. It is affected by many factors like diet and hygiene. These factors differ between different populations. Testing the population-microbiome differences using healthy samples from different countries is the first objective of this study. This data was then used in training and testing machine learning models (Random Forest and L2-logistic Regression Classifiers) for the prediction of the geographical location based on the microbiome data. Random Forest Classifier had the highest accuracy. Feature importance analysis showed that the data for Proteobacteria, Actinobacteria, and Bacteroidetes improved the Random Forest Classifier's performance. The second objective of the study was to compare the gut microbiome from healthy individuals and Coronavirus disease of 2019 (COVID-19) patients from China. COVID-19 caused lots of deaths besides an economic crisis. According to the World Health Organization (WHO), it has caused more than 227 million cases and more than 4.5 million deaths till September 16th, 2021. It was caused by severe acute respiratory syndrome coronavirus - 2 (SARS-CoV-2) which can enter the cells through the receptor for Angiotensin-converting enzyme 2 (ACE2). Proteobacteria, Actinobacteria, and Bacteroidetes had the most distinguished patterns between healthy and patient samples. Proteobacteria contain many human pathogens. Actinobacteria can cause many respiratory disorders. Bacteroidetes can regulate the expression of ACE2 receptors in mice. In conclusion, there was a correlation between being infected with SARS-CoV-2 and modifications in the gut microbiome.

Table of Contents

Chapter 1: Literature Review & Study Objectives 8

1.1. The microbiome 8

1.2. Microbiome-based population studies 9

1.3. COVID-19 9

1.4. Bioinformatics 10

1.5. Machine learning 13

1.6. Objectives 14

Chapter 2: Materials & Methods 15

2.1. Study subjects and Design 15

2.2. Downloading and installing the required software 18

2.3. Downloading and installing the required data 18

2.4. Data analysis with QIIME2 18

2.5. Data processing and development of machine learning models. 19

Chapter 3: Results and Discussion 21

Geographical Location and the microbiome 21

COVID-19 related analysis 55

Chapter 4: Conclusions and Future Perspectives 61

References 62

List of Figures

Figure 1: 16S rRNA with the hypervariable regions. Reprinted with permission of the American Thoracic Society. Copyright © 2021 American Thoracic Society. All rights reserved. Cite: Lijia Cui, Alison Morris, Laurence Huang, James M. Beck, Homer L. Twig 11

Figure 2: Flow chart depicting the whole method used in the study. 20

Figure 3:Phyla in healthy population from Italy. 22

Figure 4: Relative Abundance of the different phyla in Italy. 23

Figure 5: Phyla in healthy population from The Netherlands. 24

Figure 6: Relative Abundance of the different phyla in the Netherlands. 25

Figure 7: Phyla in healthy population from Canada. 26

Figure 8: Relative Abundance of the different phyla in Canada. 27

Figure 9: Phyla in healthy population from Sweden. 28

Figure 10: Relative Abundance of the different phyla in Sweden. 29

Figure 11: Phyla in healthy population from China. 30

Figure 12: Relative Abundance of the different Phyla in China. 31

Figure 13: Phyla in healthy population from Denmark. 32

Figure 14: Relative Abundance of the different phyla in Denmark. 33

Figure 15: Absolute abundance Analysis of healthy microbiome from 7 Egyptian samples using QIIME2. 34

Figure 16: Phyla in healthy population from Egypt. 35

Figure 17: Relative Abundance of the different phyla in Egypt. 36

Figure 18: Phyla in healthy population from Finland. 37

Figure 19: Phyla in healthy population from Finland. 38

Figure 20: Phyla in healthy population from USA. 39

Figure 21: Phyla in healthy population from USA. 40

Figure 22: Phyla in healthy population from Russia. 41

Figure 23: Phyla in healthy population from Russia. 42

Figure 24: Phyla in healthy population from Madagascar. 43

Figure 25: Phyla in healthy population from Madagascar. 44

Figure 26: Prevalence of the different phyla in the samples from the countries in the study. 45

Figure 27: Comparison between the different countries used in the study for the prevalence of Firmicutes in their healthy microbiome. 46

Figure 28: Comparison between the different countries used in the study for the prevalence of Bacteroidetes in their healthy microbiome. 47

Figure 29: Firmicutes/Bacteroidetes ratio. 48

Figure 30: Bacteroidetes/Firmicutes ratio. 49

Figure 31: Comparison between the different countries used in the study for the prevalence of Actinobacteria in their healthy microbiome. 50

Figure 32: Comparison between the different countries used in the study for the prevalence of Euryarchaeota in their healthy microbiome. 51

Figure 33: Comparison between the different countries used in the study for the prevalence of Verrucomicrobia in their healthy microbiome. 52

Figure 34: Comparison between the different countries used in the study for the prevalence of Proteobacteria in their healthy microbiome. 53

Figure 35: Comparison of the mean accuracy of Logistic Regression (LR) Classifier and random forests Classifier (RF). 54

Figure 36: Using feature importance attribute in Random Forest classifier model method. 55

Figure 37: Absolute abundance of Phyla in Covid-19 population from China using QIIME2. 56

Figure 38: Phyla in Covid-19 population from China. 57

Figure 39: Comparison between shared phyla between healthy samples and Covid-19 patients from China. 58

Figure 40: Relative Abundance of Rothia Genus in healthy samples vs COVID-19 samples from China. 59

List of Tables

Table 1:Metadata of the total dataset sets used in the study. 15
Table 2: Metadata for the combined datasets..... 21
Table 3: Random Forest vs Logistic Regression Classifiers..... 54
Table 4: Phylum vs Correlation Coefficient. 55

Glossary and Abbreviations

COVID-19	Coronavirus disease of 2019
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
QIIME	Quantitative Insights Into Microbial Ecology
WHO	World Health Organization
CD	Crohn's disease
SRA	Sequence Reads Archive.
NCBI	National Center for Biotechnology Information
TSV format	Tab delimited format
IDE	Integrated Development Environment
CSV	Comma Separated value format
OTU	Operational taxonomic unit
ASV	Amplicon sequencing variant
GI tract	Gastrointestinal tract
ACE2	Angiotensin-converting enzyme 2
USA	United States of America
OS	Operating system
LR	Logistic Regression
RF	Random Forests

Acknowledgments

I would like to thank GOD for giving me great opportunities in my life. I would like to thank both my supervisors, Dr. Asma Amleh, professor, and Dr. Ahmed Moustafa, professor at the biology department – The American University in Cairo, for giving me the chance to work under their supervision and for being great mentors in my journey at AUC. They were both mentors and role models for me.

I would like to thank Dr. Zewail -God Bless his soul- for all he did in his scientific career and for his believe in the Egyptian youth.

I would like to thank my colleagues at AUC, Zewail city, and Proteinea company for their support and encouragement.

I would like to thank people working on QIIME 2 platform for their great work. In addition, I would like to thank the researchers working on the Curated metagenomics R package for their generous and informative work. I would like to thank researchers and scientists for working on SRA and who uploaded the data from the multiple studies I used in this thesis.

Finally, I would like to thank my friends and my family, especially my parents, for believing in me and supporting me at all times with all they can. I will be in your debt forever.

Chapter 1: Literature Review & Study Objectives

1.1. The microbiome

Metagenomics is the unbiased study of the genomes in a community within an ecosystem. Two definitions are used interchangeably in the study of microbial communities. They are the microbiota and microbiome. The microbiota is a term describing the microbial communities in a population, while the microbiome is combined genetic material of the microbiota of a specific habitat. Meta-transcriptomics study the expressed genes by the members of this community. Metabolomics determines the byproducts released in the environment by the members of this community as well as its host, in case of host-associated microbial communities, to give the whole picture (Aguiar-Pulido et al., 2016).

The gastrointestinal tract in the human body has a vast and diverse community of microbes that can get to 100 trillion microorganisms (Thursby & Juge, 2017). The colon's bacterial cell density is estimated at 10^{11} to 10^{12} for each milliliter (Ley et al., 2006). The human genome has approximately 23,000 genes, while the gut microbiome has more than 3 million genes that encode thousands of proteins and metabolites (Valdes et al., 2018). The gut microbiome and its host have many intertwined interactions that result in the stability of a highly diverse and resilient community. This community has a symbiotic relationship with that host, transforming it into a "superaorganism" (Gill et al., 2006; Luckey, 1972). This superorganism can perform metabolic and immune functions (Thursby & Juge, 2017). Bacteria play a vital role in the regulation of digestion. The commensal bacteria are essential in synthesizing, extracting, and absorbing many nutrients and metabolites like bile acids, amino acids, vitamins, lipids, and short-chain fatty acids (SCFAs) (Rinninella et al., 2019). The gut microbiota prevents the invasion of bacteria by maintaining the integrity of the intestinal epithelium (Khosravi & Mazmanian, 2013). This microbiota enhances the host's immunity against pathogenic bacteria by inhibiting their growth, producing bacteriocins, and consuming available nutrients. They inhibit the colonization of the pathogenic bacteria through lots of competition processes: pH modification, nutrient metabolism, effects on the pathways for cell signaling, in addition to antimicrobial peptide secretion (Rinninella et al., 2019). They regulate the homeostasis, development, and function of both adaptive and innate immunity (Brestoff & Artis, 2013).

The microbiota in the gut includes many types of microorganisms, including yeast, and bacteria. Fungi and viruses in the GI tract constitute the gut mycobiome and the gut virome, respectively (Berg Miller et al., 2012; Dollive et al., 2012; Lopetuso et al., 2016). The different species are classified taxonomically into phyla, and each phylum is divided into classes. Each class is divided into orders, an order is divided into families, a family is divided into genera, and finally, a genus is composed of different species. A few phyla of bacteria are identified, corresponding to more than 160 species (Laterza et al., 2016). Several microbial phyla are dominant in the gut microbiota, like Firmicutes, Bacteroidetes, Actinobacteria, Proteobacteria, Verrucomicrobia, and Fusobacteria. Firmicutes and Bacteroidetes represent about 90% of the microbiota in the gut (Arumugam et al., 2011). Firmicutes, Bacteroidetes, Actinobacteria, and Proteobacteria are the predominant four major phyla (Khanna & Tosh, 2014). Bacteroidetes have two main dominant genera in the gut: *Prevotella* and *Bacteroides*. The Firmicutes includes more than 200 genera like

Bacillus, *Lactobacillus*, *Clostridium*, *Ruminococcus*, and *Enterococcus*. The genera of Clostridium are almost 95% of the Firmicutes phylum. Actinobacteria phylum is less abundant and is represented mainly through the genus of Bifidobacterium (Arumugam et al., 2011).

The gut microbiota can differ according to the anatomical region in the intestine. The GI tract is divided into different regions. These regions are different in their pH and O_2 tension, physiology, the flow rate of the digested matter, and the host secretions (Flint et al., 2012). Small and large intestines are different in their habitability for the microbiota. The small intestine has high concentrations of bile and short transit times (3 – 5 h). At the same time, the large intestine has a slower flow rate and neutral to slightly acidic pH. The gut microbes' population displays what is called a rostrocaudal gradient; the stomach and duodenum have about 10-10^3 microbes per gram, the jejunum and ileum have 10^4 – 10^7 microbes per gram, and the colon has 10^{11} - 10^{12} microbes per gram (Neish, 2009; O'Hara & Shanahan, 2006). Therefore, the large intestine has the largest microbial community, with the obligate anaerobes as the dominant microbiota (Flint et al., 2012).

1.2. Microbiome-based population studies

The study of healthy human microbiome is as important as the study of the microbiome associated with diseases. Many studies are investigating the links between the microbiome and disease. Enormous efforts are being made to understand how the microbiome changes with genetics, host lifestyle, age, medication, nutrition, and environment (Blaser et al., 2008; Blekhman et al., 2015; Zhan Gao et al., 2008; Turnbaugh et al., 2006; A Zhernakova et al., 2016). The study of the microbiome of different human populations is important. Variations in the microbiome based on the population depends on many factors. One example is the population of Netherlands, where people consume less antibiotics and more dairy products in comparison to other populations from Europe (Alexandra Zhernakova et al., 2016). Therefore, the population-based microbiome profile from healthy individuals should be considered for the identification of the different geographical and ethnic microbiome biomarkers, before studying other factors associated with a particular disease (Gupta et al., 2017). The geographical location gives a plethora of environmental, genetic, and cultural factors to which degree the microbiome is shaped by each of these factors remains to be explored.

1.3. COVID-19

The COVID-19 pandemic has changed the lives of most people around the world. It has caused many losses in human lives besides an economic crisis. According to the WHO, more than 227 million cases of COVID-19 were recorded and resulted in more than 4.5 million deaths till September 16th of this year, 2021. It is caused by the severe acute respiratory syndrome coronavirus - 2 (SARS-CoV-2). Although many people have been infected by the virus, most cases are mild. Severe cases can result in hospitalization, respiratory system failure, or death (Onder et al., 2020). In the early reports from Wuhan, there were symptoms in the gastrointestinal tract in 2% to 10% of the patients with COVID-19, like diarrhea, but in a meta-analysis, results showed that the GI symptoms were in at least 20% of the COVID-19 patients (Chen et al., 2020; Cheung et al., 2020; Huang et al., 2020; Liang et al., 2020). It is mainly transmitted through either respiratory droplets or aerosols from the breath of the infected individuals and inhaled by another

person. Other ways of transmission were reported, like fecal-oral transmission (Gallardo-Escárate et al., 2021; Y. Xu et al., 2020) and transmission by fomite (Sia et al., 2020). Many studies have found that SARS-CoV-2 can be detected in the stool samples and anal swabs in about 50% of the COVID-19 patients, which indicates that virus replication and activity can occur out of the pulmonary system (Wölfel et al., 2020; Y. Xu et al., 2020).

The bacteria in the gut can control the infectivity of human viruses that can be transmitted by fomite through improving their thermostability (Berger et al., 2017), improving the environmental stability (Robinson et al., 2014), in addition to encouraging the diversity and fitness of viruses (Erickson et al., 2018). The interaction between a virus and a bacterium was observed in infections in the upper respiratory tract like influenza A (Tashiro et al., 1987) and the infection by oral human papillomavirus (Pavlova et al., 2019). In the human microbiome, prevalent bacteria were found to alter the human glycocalyx to control the capability of binding SARS-CoV-2 to the host cells (Martino et al., 2020).

Angiotensin-converting enzyme 2 (ACE2) receptor is highly expressed in the respiratory system and the GI tract (Shang et al., 2020; J. Wang et al., 2020; Xiao et al., 2020). SARS-CoV-2 enters the host through (ACE2) receptor (Shang et al., 2020). It is important in controlling both the inflammation and microbial ecology in the gut (Hashimoto et al., 2012). The microbiota in the gut is dynamic. It is regulated by infection with viruses to allow for stimulatory or suppressive response (N. Li et al., 2019). Many studies have shown that viral infections in the respiratory system can be associated with changes in the gut microbiota, which results in predisposing patients to infections with bacteria (Hanada et al., 2018; Yildiz et al., 2018). Alterations in the phyla of Bacteroidetes and Firmicutes were found to be associated with comorbidities associated with severe cases of COVID-19 (Emoto et al., 2016; Ley et al., 2006; Turnbaugh et al., 2006; Yang et al., 2015), and both phyla were found to regulate ACE2 expression in mice (Geva-Zatorsky et al., 2017). A study from Hong Kong (Zuo et al., 2020) concluded that approaches that aim at altering the microbial communities in the intestine might reduce the severity of COVID-19. They found that there were constant alterations in the fecal microbiome during the hospitalization time. In addition, they found these alterations were related to the levels of SARS-CoV-2 in the fecal samples and the disease's severity.

1.4. Bioinformatics

In the last two decades, technology advanced at a swift rate in the biological sciences. The tremendous and quick advances in DNA sequencing significantly impacted biological research. This led to many applications in medicine, plant science and agriculture, and environmental sciences. One of the significant fields impacted by these advances in microbiology. The classical way of researching microbiology was to isolate the samples and culture the microbial species from the sample. This approach was unsuccessful with most of the samples as many species, primarily anaerobic species, cannot be grown in the lab (Bradshaw et al., 1996). The advances in DNA sequencing and the associated advancements in bioinformatics software to analyze these data lead to a significant change in this approach to microbiology. Now, there is no need to culture the microbes in the lab to do the analysis or taxonomic analysis. Using Marker genes like the internal-transcribed-spacer regions in fungi, 16S rRNA genes for bacteria and

archaea, and 18S rRNA genes in eukaryotes, the microbiota can be taxonomically and phylogenetically profiled with varying degrees. The integration of other types of data such as metaproteome (Verberkmoes et al., 2009), meta-transcriptome (Barr et al., 2018), and metabolic profiles (Kapono et al., 2018) will increase and enhance the specificity of the analysis.

The study and characterization of the microbiome are done through targeted sequencing of conserved genes like the hypervariable regions in the 16S ribosomal RNA gene (Huse et al., 2012) figure (1). 16S rRNA sequencing is one of the most used housekeeping genetic markers to study the microbiome. It is present in almost all bacterial species, existing as a multigene family or operons. Its function has not changed over time. It is large enough to be used in bioinformatics analysis (Patel, 2001). Between 1980 and now, there has been an explosion in the number of recognized taxa. This explosion directly relates to how easily 16S rRNA sequencing can be performed compared to manipulating DNA-DNA hybridization experiments. DNA-DNA hybridization is considered the golden standard for identifying and proposing a new species and confidently assigning it to a suitable taxonomic unit (Janda & Abbott, 2007). The improvement and the costs are getting lower over the years for sequencing nucleic acids. These are two main reasons for the spread and utilization of sequencing-based analysis.

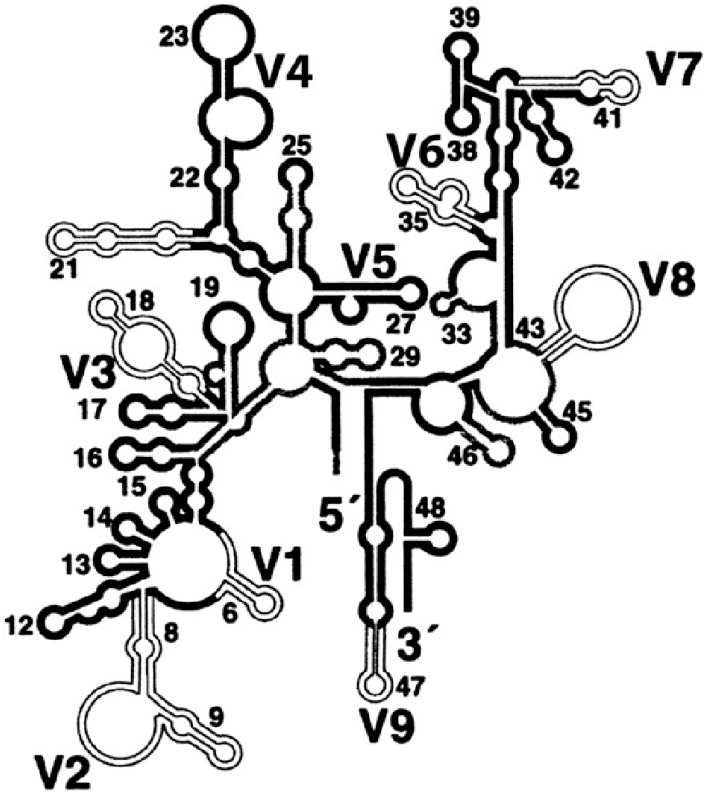

Figure 1: 16S rRNA with the hypervariable regions. Reprinted with permission of the American Thoracic Society. Copyright © 2021 American Thoracic Society. All rights reserved. Cite: Lijia Cui, Alison Morris, Laurence Huang, James M. Beck, Homer L. Twig

Two standard classification methods are used for sequences from a microbial community: operational taxonomic unit (OTU) and Amplicon sequencing variant (ASV). Clustering with OTU was proposed based on that related or similar organisms have similar gene sequences that can be targeted. Sequencing errors are rare and may not have any contributions or just trivial contributions to the consensus sequence of these clusters (Blaxter et al., 2005). There are three main methods to generate OTUs from the data. Clusters are usually generated using 97% sequence identity as similarity threshold. This may lead to grouping many similar species into the same OTU, and so their individuality may be lost. Some researchers tried to use higher similarity thresholds like 100% to lower the risk of diversity loss due to clustering, but this created an increased risk of assigning new species to sequencing errors and so get false diversity (Kunin et al., 2010). OTU clustering has the advantage of minimizing the influence of the sequencing errors

in the reads' pool through clustering the similar sequences into the abstracted consensus sequence. ASV determines which exact sequences were read and how many times every sequence was read. Then the data is combined with a model for error in the sequencing run. This would allow the researcher to compare similar reads and determine the probability that a given read with a specific frequency is not due to randomness or sequencing error (Caporaso et al., 2010).

Many different tools and pipelines were developed to analyze metagenomic data (Kim et al., 2013). Many problem-solving environments allow the user to use user-friendly workbenches and develop analysis pipelines that are flexible and easy from available software tools (Gallopoulos et al., 1994). Some of these platforms are Galaxy (Goecks et al., 2010), Mothur (Schloss et al., 2009), Pathoscope (Hong et al., 2014), and Quantitative Insights Into Microbial Ecology (QIIME) (Caporaso et al., 2010). Galaxy provides the user with a framework for genomic analysis and the needed tools and file formats for various steps in the pipeline. It allows the remote execution of jobs, halting and repeating individual steps, and permitting intermediate processes and results' inspection. Galaxy's main disadvantages are the high computational and storage requirements. Mothur supplies the user with a functionally accessible and extensible package through a domain-specific language (Aguiar-Pulido et al., 2016). Pathoscope gives the user a specialized pipeline to identify bacterial strains from the raw sequences and generate statistics like percentages, protein products, and gene locations. It is open-source and able to accommodate any user, tool, or developer. QIIME allows the user to integrate scripts to analyze raw microbial DNA samples, for example, taxonomic classification using marker genes like 16S rRNA. QIIME also enables the construction of flexible pipelines.

The platform used in this study is QIIME2. QIIME2 is a system developed with a plugin architecture, and it allows the contributed functionality by a third party (Bolyen et al., 2019). QIIME2 provides a large set of plugins ranging from the latest generation of tools for sequence quality control like DADA2 (Callahan et al., 2016) and Deblur (Amir et al., 2017). In addition to plugins for taxonomic assignment (Bokulich, Kaehler, et al., 2018) and phylogenetic insertion (Janssen et al., 2018). Recently, there were new plugin releases that provide support for metabolomics analysis and shotgun metagenomics data analysis like q2-cscs (Sedio et al., 2018), q2-shogun (Hillmann et al., 2018), q2-metaphlan2 (Truong et al., 2015), q2-metabolomics (M. Wang et al., 2016), and q2-picrust2 (Langille et al., 2013). In addition, QIIME2 has plugins for machine learning (Bokulich, Dillon, et al., 2018). QIIME2 has excellent potential for becoming a multidimensional platform for data science and being easily and rapidly used to analyze different microbiome features (Bolyen et al., 2019). QIIME2 also provides interactive visualization tools. The visualization results from the pipeline can be viewed using (https://view.qiime2.org). It is a free and unique service that allows different users to share and manipulate the results securely without installing QIIME2 (Bolyen et al., 2019).

CuratedMetagenomicData R package (Pasolli et al., 2017) is the source of most of the data in the study. The CuratedMetagenomicData provides standardized and curated human microbiome data. It is distributed through the Bioconductor ExperimentHub platform. It allows the user to access human microbiome data that is processed uniformly. It includes the taxonomic abundance of bacteria, viruses, archaea, and fungi. Also, it provides the user with the metadata for

each sample and the metabolic functional profiles. The resources for the data are available to users with the very low knowledge of bioinformatics and can be integrated easily with the R programming environment (Team, 2020). R programming language is considered one of the easiest and most flexible programming languages. Therefore, CuratedMetagenomicData R package provides a great opportunity for clinicians, biologists, epidemiology statisticians to do novel analyses and develop methodologies that can integrate the results in novel ways.

1.5. Machine learning

Machine learning is a technology with great potential. It has been used in many different fields, from engineering to medicine. It has been popular in microbiome research as it can account for the interpersonal variations in the microbiome and the disease's ecology. Machine learning can be used with different types of data. The relative abundance of bacterial populations can be considered for each population while considering the other populations (Topçuoğlu et al., 2020). Machine learning models enhance the research about the differences in the available data and allow predictions about the new data. Machine learning models were developed to diagnose a variety of diseases like liver cirrhosis (Qin et al., 2014), inflammatory bowel diseases (Mossotto et al., 2017), type 2 diabetes (Walters et al., 2014), colorectal cancer (Baxter, Koumpouras, et al., 2016; Baxter, Ruffin, et al., 2016; Dadkhah et al., 2019; Zackular et al., 2014; Zeller et al., 2014), obesity (Sze & Schloss, 2016; Walters et al., 2014), and skin cancer (Soenksen et al., 2021).

In a recent study (Topçuoğlu et al., 2020), multiple machine learning models were trained on 16S rRNA sequence data to predict the presence or absence of colonic neoplasia. The study aimed to assess the predictive performance, the time needed for training the models, and the interpretability of the results to show the effects of model selection. Both the random forest model and the L2-logistic regression model gave the best results. The random forest model had the best performance but a long time for training and hard interpretability. L2-logistic regression model followed random forest in the performance, but it had shorter training time and better integrability.

Logistic regression is a statistical method that learns a specific model that can predict the outcome of a binary variable from one or multiple response variables that are continuous or categorical (Hoffman, 2019). It is a type of supervised learning; it both trains and evaluates the models based on data input supplemented with labels to indicate the outcomes for the given information. There are many supervised learning approaches, including statistical classification and regression analysis (Marcos-Zambrano et al., 2021). Logistic regression classification can be used with multivariable classification problems by turning it from being a binary classifier into a multinomial logistic regression classifier. This can be done through different techniques. One technique is to divide the multi-class classification problem into many more minor binary classification problems, known as one-vs-rest (Rifkin & Klautau, 2004). Another approach is to modify the logistic regression classifier to support the classification and prediction of multiple class labels through using a multinomial probability distribution. Multinomial probability distribution allows the logistic regression classifier to define multi-class probabilities (Starkweather & Moske, 2011). Logistic regression was used in the research of bacterial vaginosis to classify and predict the microbial signatures (Beck & Foster, 2015). It was used in an extensive study of 300 biomarkers and the relative gene abundance from the

microbiomes of individuals from China. It could achieve outstanding performance and accuracy. This study could show that the microbiome biomarkers from the gut can distinguish abnormal cases from the controls with high specificity (H. Wu et al., 2018).

Random forest is a type of ensemble learning which combines multiple classifiers to get better performance than using a single classifier. The simple models in the random forests are decision trees. It uses bootstrapping on the dataset to derive individual decision trees. The final result is the majority voting of the single decision trees (Marcos-Zambrano et al., 2021). It has been used to classify pediatric patients with Crohn's disease (CD) (Douglas et al., 2018). It was used to find biomarkers (Koohi-Moghadam et al., 2019; Thomas et al., 2019; Wirbel et al., 2019) and analyze host-microbial signatures to detect the contamination of feces from environmental samples (Roguet et al., 2018).

Determining the essential features in the datasets used in machine learning is an important step for analyzing the data and increasing the model's accuracy. Feature selection uses correlation analysis to analyze the different features in the data and determine the best feature that should be used to increase the accuracy of the model. Random forest classifier has an attribute called feature importance that allows for analyzing the importance for each column or variable in the dataset. It measures the information gain or impurity for each node in a decision tree. As the random forest model produces multiple decision trees, it calculates the average information gain or impurity over all the trees in the model based on each variable. The most important feature is the one with the highest decrease in impurity. However, this method has advantages like the speed of computation. It has some disadvantages, like preferring numerical features over categorical features. In addition to that, it may neglect one feature over the other in the case of correlated data (Pedregosa et al., 2011). There are different methods for determining feature importance like Permutation importance and calculating feature importance using Shapley values. Permutation importance shuffles each feature in a dataset randomly then computes the performance of the model. Then the features that improve the performance of the models are declared the most important (Pedregosa et al., 2011). The third method relies on calculating Shapley values from game theory to estimate how each feature affects the predictivity of the model (Lundberg & Lee, 2017). It has better figures as it provides for examining each feature according to the different categories of the output.

1.6. Objectives

The objectives of the study were to

- Test the correlation between the geographical location and the gut microbiome from healthy individuals and develop a predictive machine learning model that can use the data as biomarkers.

- Compare the gut microbiome from healthy individuals and COVID-19 patients from China.

Chapter 2: Materials & Methods

2.1. Study subjects and Design

The data and metadata for 16 studies with 1445 healthy subjects from Italy, Netherlands, Canada, Sweden, China, Denmark, USA, Finland, China, Madagascar, and Russia was extracted from the curated metadata R package (Pasolli et al., 2017). In addition, the relative abundance data for healthy subjects from Egypt and the relative abundance data for COVID-19 patients from China was generated directly after analyzing 16S rRNA Fastq samples using QIIME2, as illustrated below. All healthy individuals in the study were naïve to antibiotic therapy. The subjects' age was as follows and divided into categories: 1 – 12 years old were denoted Child, 12 – 19 years old were denoted School age, 20 – 65 years old were denoted Adult, and 65 years old < were denoted Seniors.

Table 1:Metadata of all dataset sets used in the study.

Country	**Dataset**	**PMID**	**Subjects' age category**	**Sex F/M/NA**	**Health state**
Italy	1) ThomasAM_2018b.metaphlan_bugs_list.stool (Thomas et al., 2019)	Unpublished	• Adult	NA	Healthy
The Netherlands	2) LiSS_2016.metaphlan_bugs_list.stool (S. S. Li et al., 2016)	27126044	• Adult	0/5	Healthy
	3) SchirmerM_2016.metaphlan_bugs_list.stool (Schirmer et al., 2016)	27984736	• School age • Adult • Senior	265/200	Healthy
Canada	4) RaymondF_2016.metaphlan_bugs_list.stool (Raymond et al., 2016)	26359913	• Adult	21/15	Healthy

Sweden	5) BackhedF_2015.metaphlan_bugs_list.stool (Bäckhed et al., 2015)	25974306	• Child • Adult	40/30	Healthy
China	6) Temporal dynamics of human respiratory and gut microbiomes during the course of COVID-19 in adults (R. Xu et al., 2020)	BioProject ID PRJNA639286	• NA	NA	COVID-19 patients
	7) LiJ_2017.metaphlan_bugs_list.stool (Jing Li et al., 2017)	28143587	• Adult	NA	Healthy
	8) JieZ_2017.metaphlan_bugs_list.stool (Jie et al., 2017)	29018189	• Adult • Senior • Na	101/69/1	Healthy
	9) LiJ_2014.metaphlan_bugs_list.stool (Junhua Li et al., 2014)	24997786	• Adult	NA	Healthy
	10) YeZ_2018.metaphlan_bugs_list.stool (Ye et al., 2018)	30077182	• Adult • Senior	11/34	Healthy
Denmark	11) HansenLBS_2018.metaphlan_bugs_list.	30425247	• Adult • Senior	116/92	Healthy

	stool (Hansen et al., 2018)				
Egypt	12) A comparative study of the gut microbiome in Egyptian patients with Type I and Type II diabetes (Radwan et al., 2020)	BioProject ID PRJNA629382	• Adult	NA	Healthy
Finland	13) VatanenT_2016.metaphlan_bugs_list.stool (Vatanen et al., 2016)	27259157	• Child	13/22	Healthy
USA	14) HanniganGD_2017.metaphlan_bugs_list.stool (Hannigan et al., 2017)	Unpublished	• Adult	NA	Healthy
	15) HMP_2012.metaphlan_bugs_list.stool (Huttenhower et al., 2012)	22699609	• School age • Adult	65/82	Healthy
	16) Obregon-TitoAJ_2015.metaphlan_bugs_list.stool (Obregon-Tito et al., 2015)	25807110	• Child • Adult	8/14	Healthy
Russia	17) VatanenT_2016.metaphlan_bugs_list.st	27259157	• Child	11/13	Healthy

	ool (Vatanen et al., 2016)				
Madagascar	18) PasolliE_2018.metaphlan_bugs_list.stool (Pasolli et al., 2017)	Unpublished	• School age • Adult • Senior	46/51/15	Healthy

2.2. Downloading and installing the required software

Linux Ubuntu 18.04 LTS operating system (OS) was downloaded and installed beside windows 10. Then Anaconda (Inc., 2020) was downloaded and installed on Ubuntu 18.04 LTS to install and use QIIME2 on Ubuntu 18.04 LTS. Then Kemi add-on was installed on google sheets to check and confirm that metadata files can be used with QIIME2. SRA toolkit (the SRA Toolkit Development Team., n.d.) was downloaded and installed on Ubuntu 18.04 LTS to be able to download the Fastq files from the Sequence Read Archive database (SRA) on National Center for Biotechnology Information (NCBI) to the local computer. Then R and the R language integrated development environment (IDE); R studio (Team, 2020) were downloaded and installed on Ubuntu 18.04 LTS and Windows 10. Curated Metagenomic Data R package was then installed on R on Windows 10. Curated Metagenomic Data allows the user to use and share preprocessed human microbiome data that is uniformly processed. These data include archaeal, viral, bacterial, and fungal taxonomic abundance, besides functional quantitative profiles and standardized metadata that is standardized (Pasolli et al., 2017).

2.3. Downloading and installing the required data

From (SRA) on NCBI (Leinonen et al., 2010), a text file with sample IDs for the Egyptian dataset from the study (Salah et al., 2019) was downloaded and used to download the 16S rRNA for each sample (BioProject ID PRJNA629382) using SRA toolkit software - Ubuntu 18.04 LTS edition. In addition, 48 COVID-19 samples for the Chinese dataset (R. Xu et al., 2020) were downloaded the same way (BioProject ID PRJNA639286).

2.4. Data analysis with QIIME2

To start the analysis on QIIME2, we imported the Fastq files for the using a file manifest which a tab-delimited format (tab-separated values, tsv) text file. This manifest file contains two or three columns according to the type of the reads: the sample id column, the paths to the forward reads, and the paths to the reverse reads for paired-end reads and the sample id column, and the paths to the reads for Single-end reads. To automate the process of importing the data into QIIME2, we developed and executed a code using R studio (Team, 2020) on Linux Ubuntu 18.04 LTS. After importing the data into QIIME2 version 2020.11, three steps were done. For the Egyptian samples, the samples were already demultiplexed paired-end sequences. So, they were joined and denoised according to the quality scores using DADA2 (Callahan et al., 2016) plugin with a truncation length of 160, and bases with quality less than 30 were removed as

mentioned in the parent paper (Salah et al., 2019). For the Chinese COVID-19 samples (R. Xu et al., 2020), The samples were also demultiplexed but single end sequences, and so they were denoised directly according to the quality scores using DADA2 (Callahan et al., 2016) plugin with a truncation length of 160 and bases with quality less than 30 were removed. For both studies, Taxonomic analysis using OTUs was done using gg-13-8-99-515-806-nb-classifier trained on the green genes database (DeSantis et al., 2006). The results were as in figure (3) and figure (35) based on the absolute abundance of the phyla. Then the absolute abundance was turned into the relative abundance using QIIME2 feature-table plugin that was also used for OTU picking.

2.5. Data processing and development of machine learning models.

Then, the relative abundance of the following microbial phyla (Firmicutes, Actinobacteria, Bacteroidetes, Euryarchaeota, Verrucomicrobia, and Proteobacteria) for all the healthy samples (1452 samples) -including Egyptian samples- was added to a comma-separated format file (CSV) with the mentioned countries. These phyla were specifically chosen because they included the four major phyla in addition to other shared phyla that ranged from bacteria to archaea.

As mentioned above, the data was stored in a CSV file and then randomized -to avoid the bias in the models- through assigning a random ID for each sample then sorting the samples based on this ID from lowest to highest. Then this dataset was used for training machine learning models. The random forest classification model and L2-logistic regression classification model were developed and trained based on the dataset, with the data divided randomly into 75% training set and 25% testing set.

The random forest classification model allows for determining the essential features in the dataset. These critical features are the most relied on in the classification. The feature importance was determined using three methods; feature importance attribute in random forest classifiers from sci-kit learn library (Pedregosa et al., 2011). The feature importance attribute from the random forest classifier allows for analyzing the importance of each column or variable in the dataset. It measures the information gain or impurity for each node in a decision tree. As the random forest model produces multiple decision trees, it calculates the average information gain or impurity over all the trees in the model based on each variable. The most important feature is the one with the highest decrease in impurity. Grid search (LaValle et al., 2004) was used to increase the accuracy of the random forest model.

All the codes including the codes for the models can be accessed and viewed on this GitHub Repository.

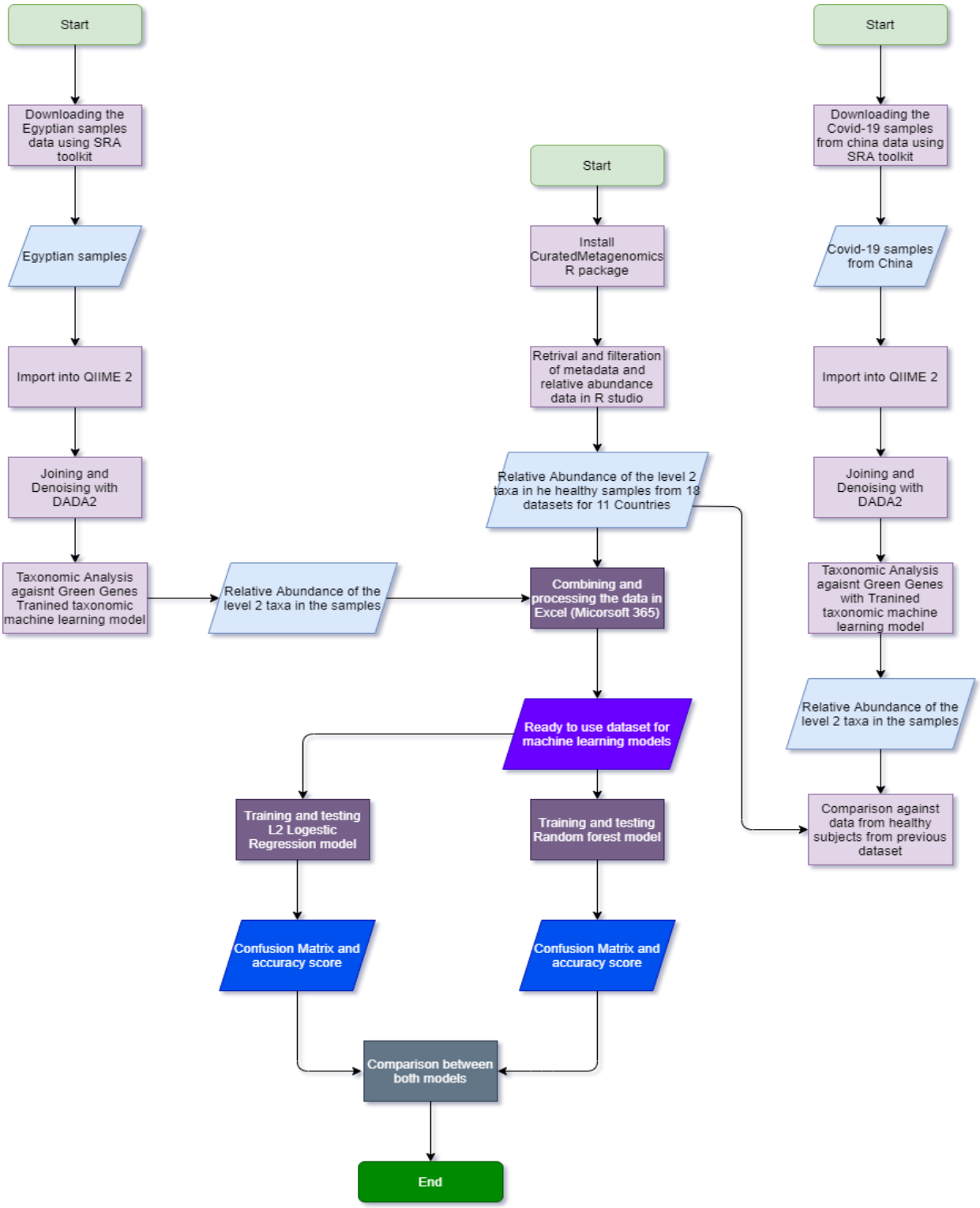

Figure 2: Flow chart depicting the whole method used in the study.

Chapter 3: Results and Discussion

Geographical Location and the microbiome

The 16S amplicon sequence datasets were used from 11 countries, with samples ranging from 7 (Egypt) to 470 (Netherlands) per country (median = 70 samples per country and a total of 1452 samples) (Table 2). This data was then combined into one dataset for training and testing the machine learning models.

Table 2: Metadata for the combined datasets.

Country	**Population by July 2020**	**No. of samples**	**Age range**
Italy	60,003,471	28	NA
The Netherlands	17,280,397	470	12 – 65 and over 65 years old
Canada	38,005,238	36	20 – 65
Sweden	10,360,869	70	1-12 and 20 – 65 years old
China	1,394,015,977	268	20 – 65 and over 65 years old
Denmark	5825337	208	20 – 65 and over 65 years old
Egypt	104,124,440	7	20 – 65 years old
Finland	5,571,665	35	1- 12 years old
USA	332,639,102	194	1 – 65 years old
Russia	141,722,205	24	1 - 12 years old
Madagascar	26,955,737	112	12 – 65 and over 65 years old
Total	1452		

The relative abundance of the highest phyla for each dataset was compared to their relative abundance in the whole dataset (Figures 3 through 25). In addition, stacked bar plots represent the relative abundance in each sample according to the dataset. The observed results indicate that although countries differ in their geographical location, they can be categorized according to the relative abundance of the different phyla in their microbiome. In addition, upon this categorization, similar patterns of phyla can be observed.

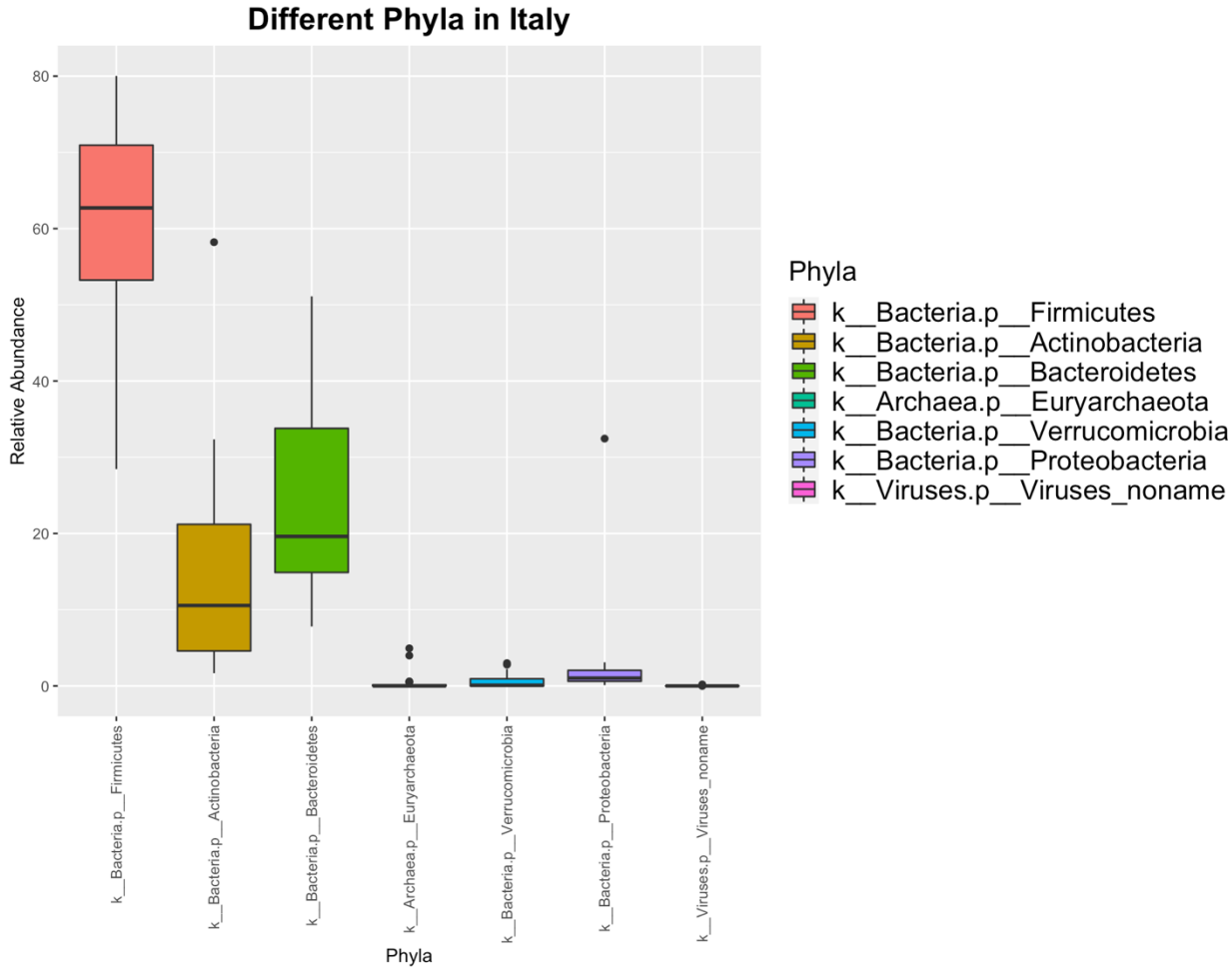

Figure 3:Phyla in healthy population from Italy.

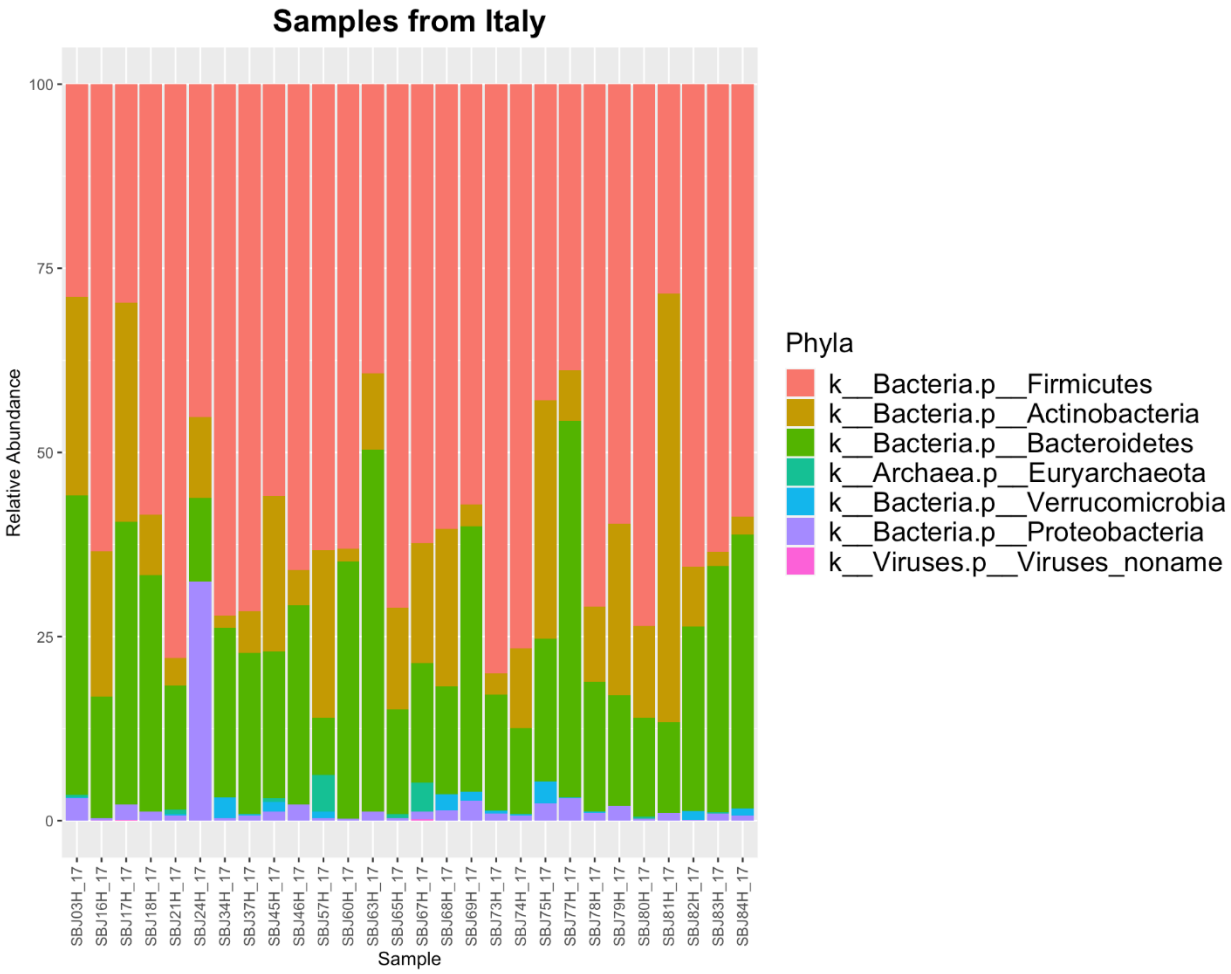

Figure 4: Relative Abundance of the different phyla in Italy.

The relative abundance of seven microbial phyla in healthy samples from 28 adults from Italy (Thomas et al., 2019) was checked (figures 3 and 4). Firmicutes represent the most abundant taxon in the six phyla, with Bacteroidetes and Actinobacteria as second and third abundant phylum, respectively.

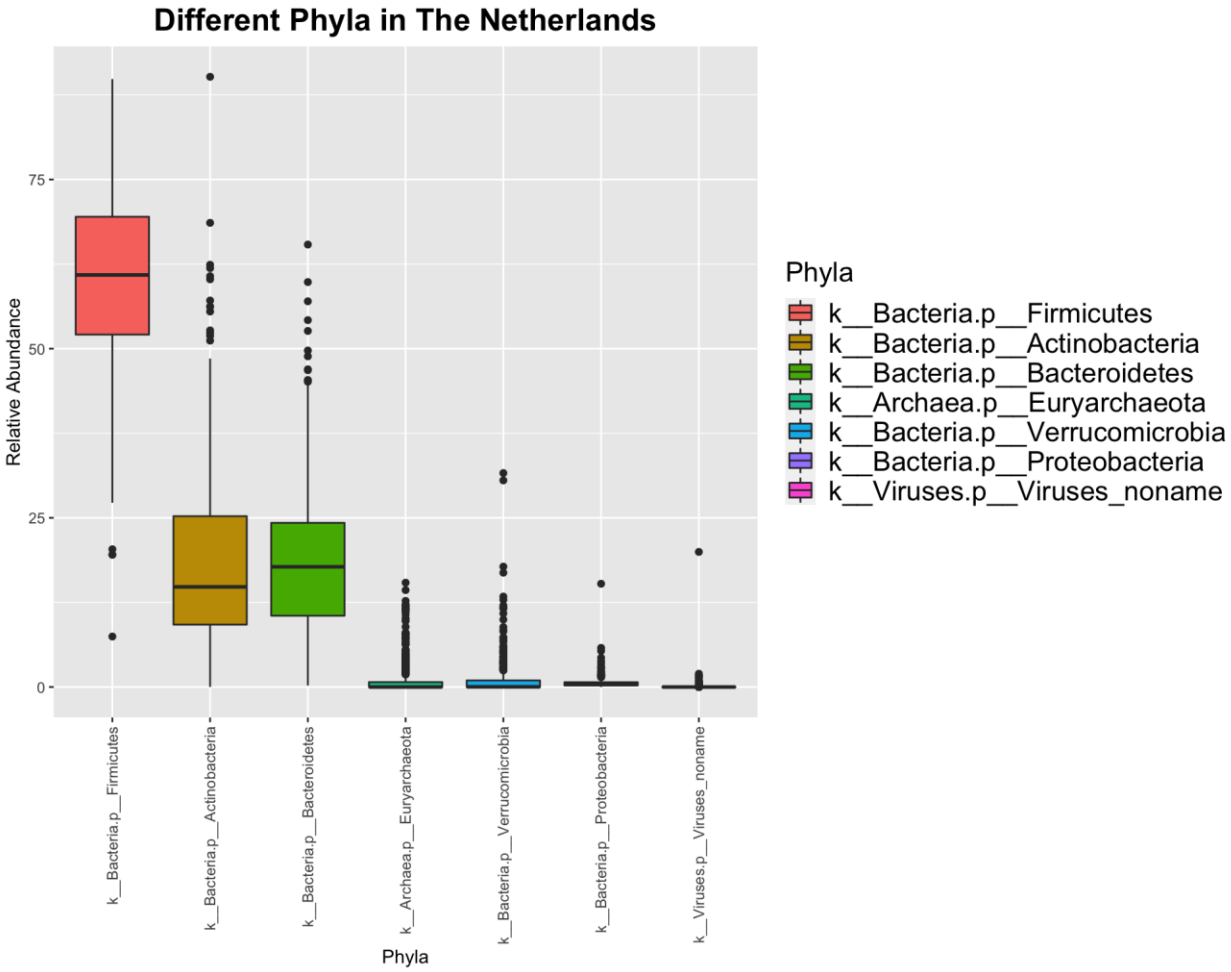

Figure 5: Phyla in healthy population from The Netherlands.

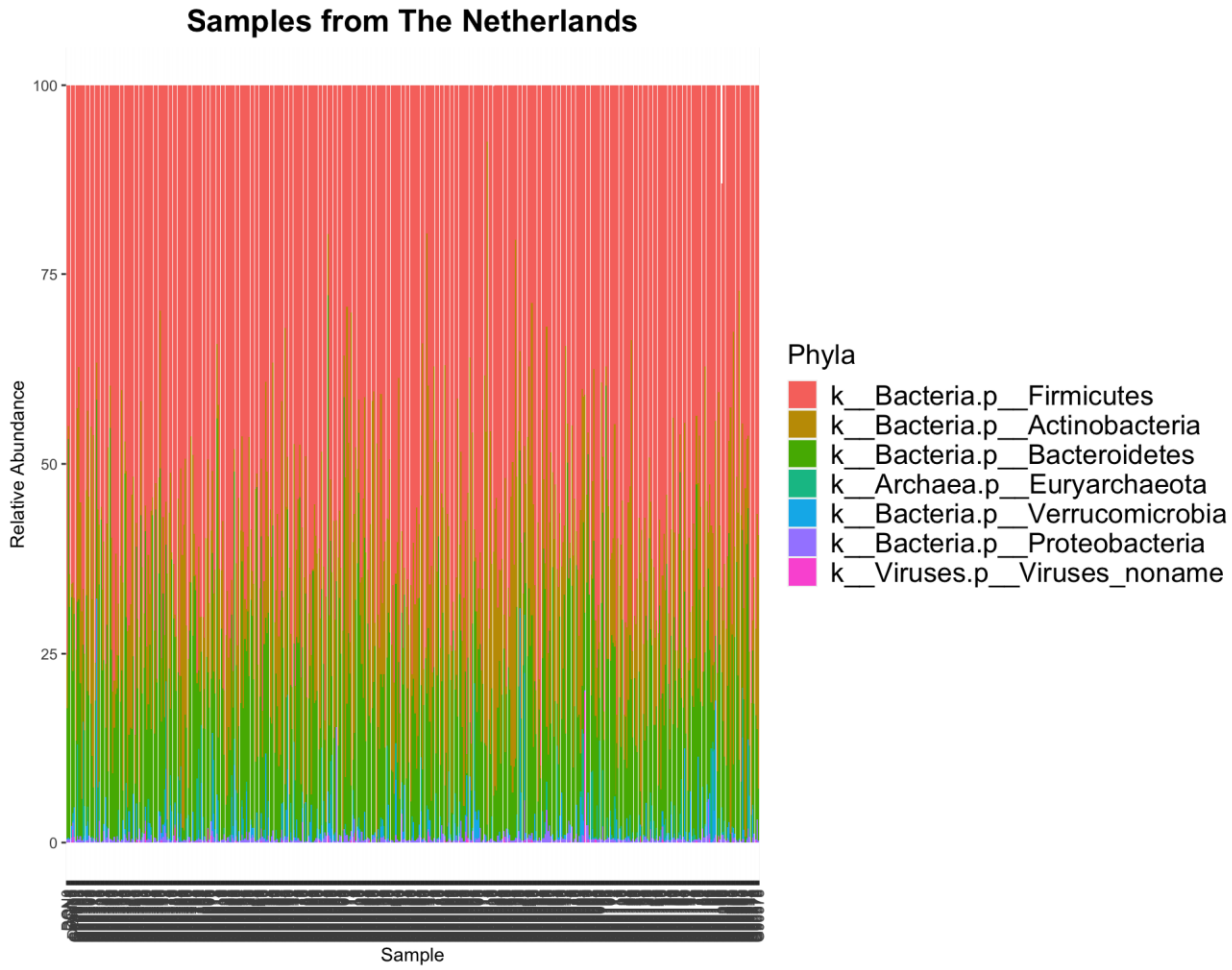

Figure 6: Relative Abundance of the different phyla in the Netherlands.

The data from two studies from Netherlands (S. S. Li et al., 2016; Schirmer et al., 2016) was retrieved (figures 5 and 6). The data included the relative abundance and metadata of 470 healthy individuals -205 males and 265 females in three age categories (School age, Adult, and Seniors). It followed a similar pattern to the data from Italy. Firmicutes had the highest abundance, with a median of almost 60%. Bacteroidetes and Actinobacteria had identical abundance. Proteobacteria, usually one of the four major phyla, had a very low abundance near 0%.

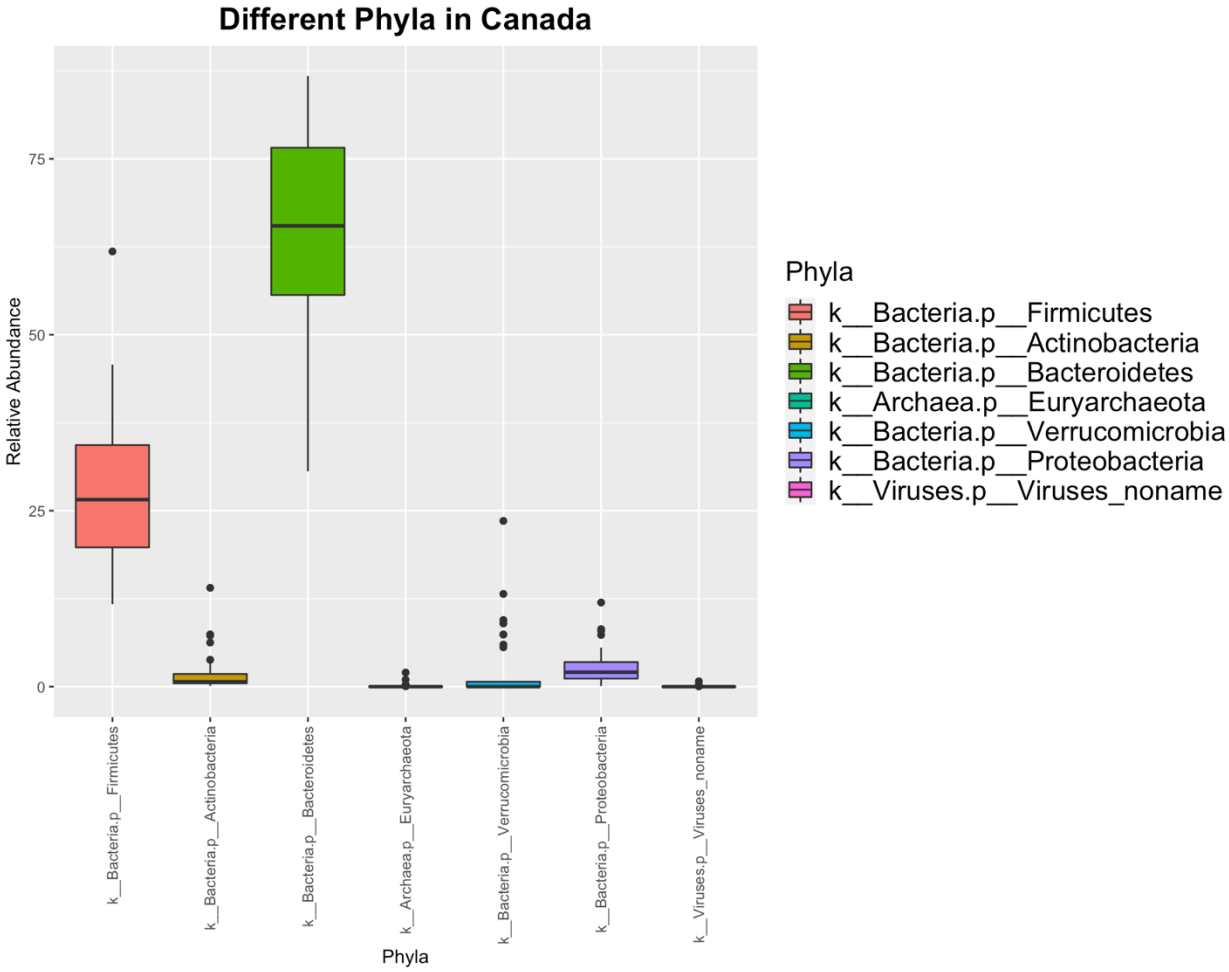

Figure 7: Phyla in healthy population from Canada.

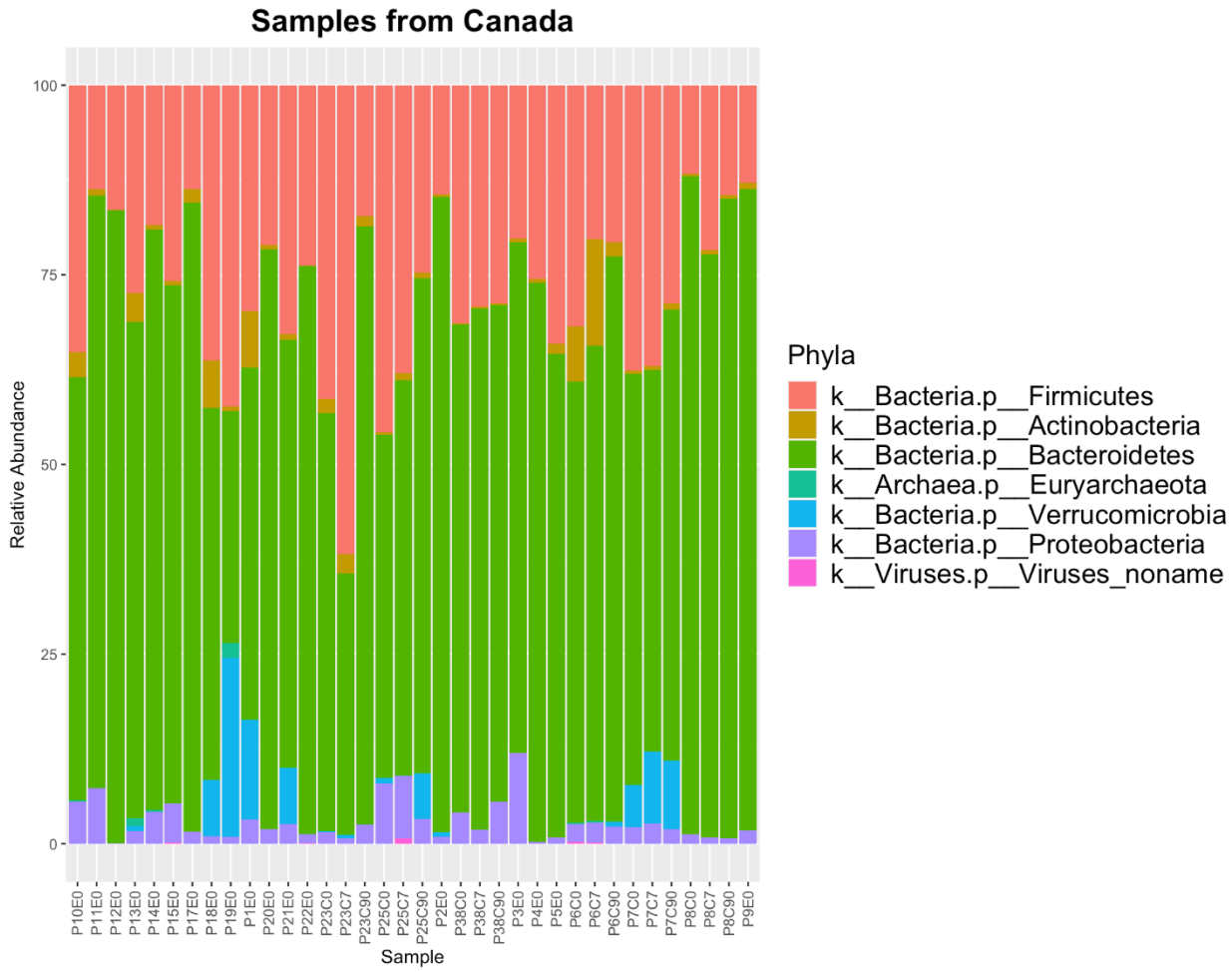

Figure 8: Relative Abundance of the different phyla in Canada.

The data analyzed was for 36 adult healthy individuals -21 females and 15 males- (Raymond et al., 2016). These individuals had Bacteroidetes as their most abundant phylum, with a median of almost 65% (Figures 7 and 8). Firmicutes had the second-highest abundance,

with a median of 28%. The other three phyla (Actinobacteria, Euryarchaeota, and Proteobacteria) had low relative abundance near to 0%.

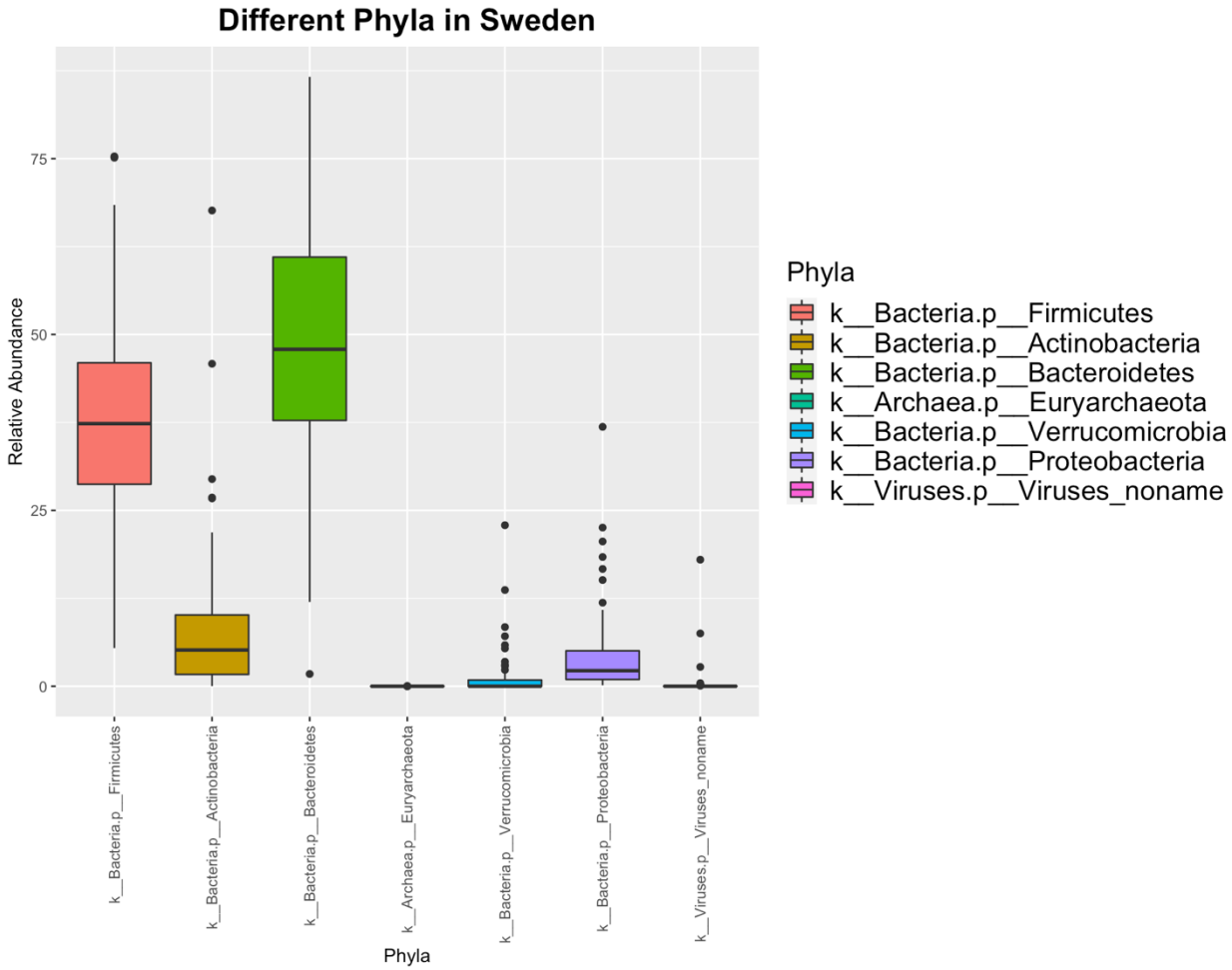

Figure 9: Phyla in healthy population from Sweden.

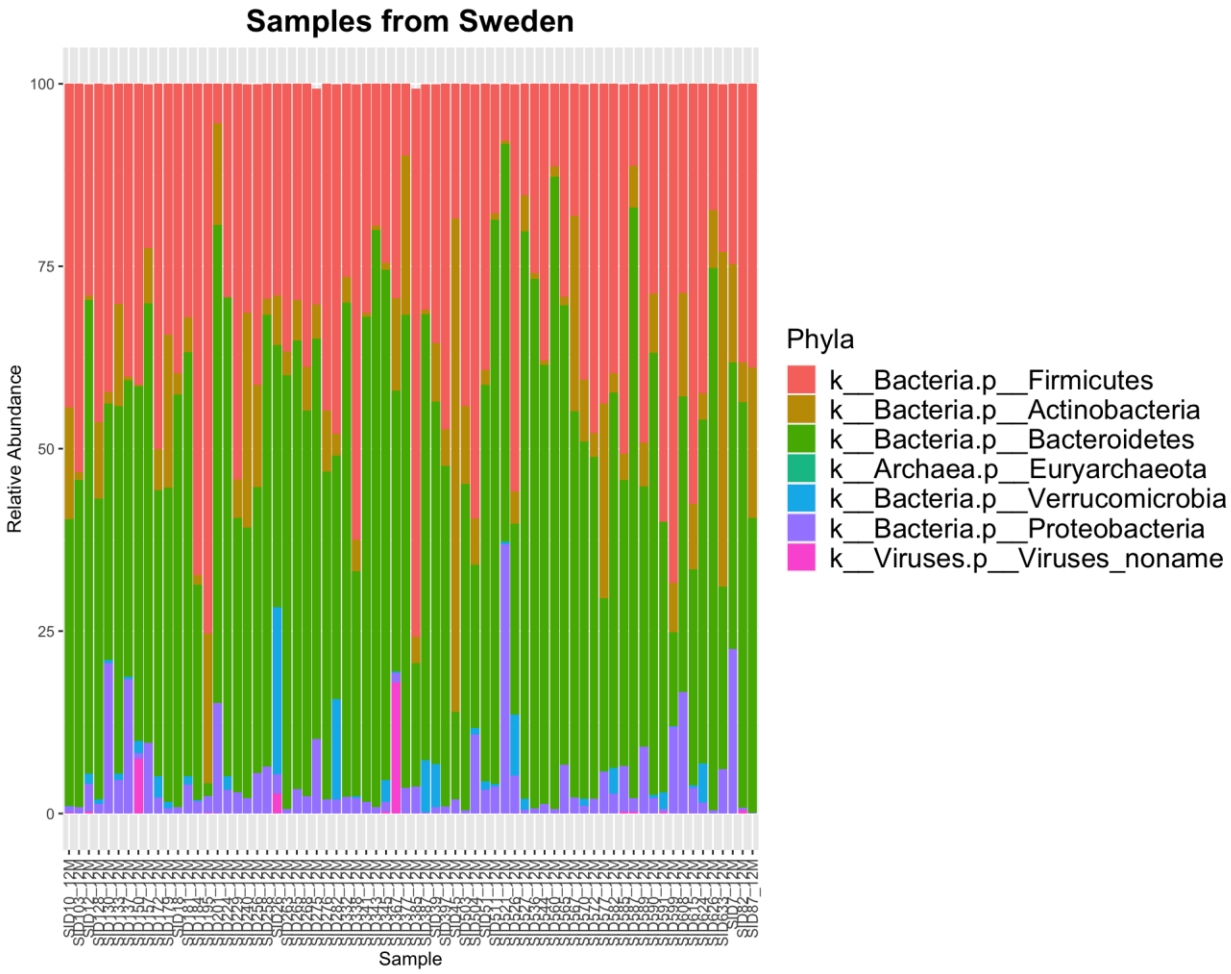

Figure 10: Relative Abundance of the different phyla in Sweden.

The relative abundance data and metadata of 70 healthy individuals (adults and children) — 40 females and 30 males —was obtained from a study in 2015 (Bäckhed et al., 2015) (figures 9 and 10). The Bacteroidetes had the highest relative abundance in the samples, with a median of 49%. Firmicutes had the second-highest relative abundance in the samples, with a median of 37.5%. Actinobacteria came third, and proteobacteria had the lowest relative abundance in the four phyla.

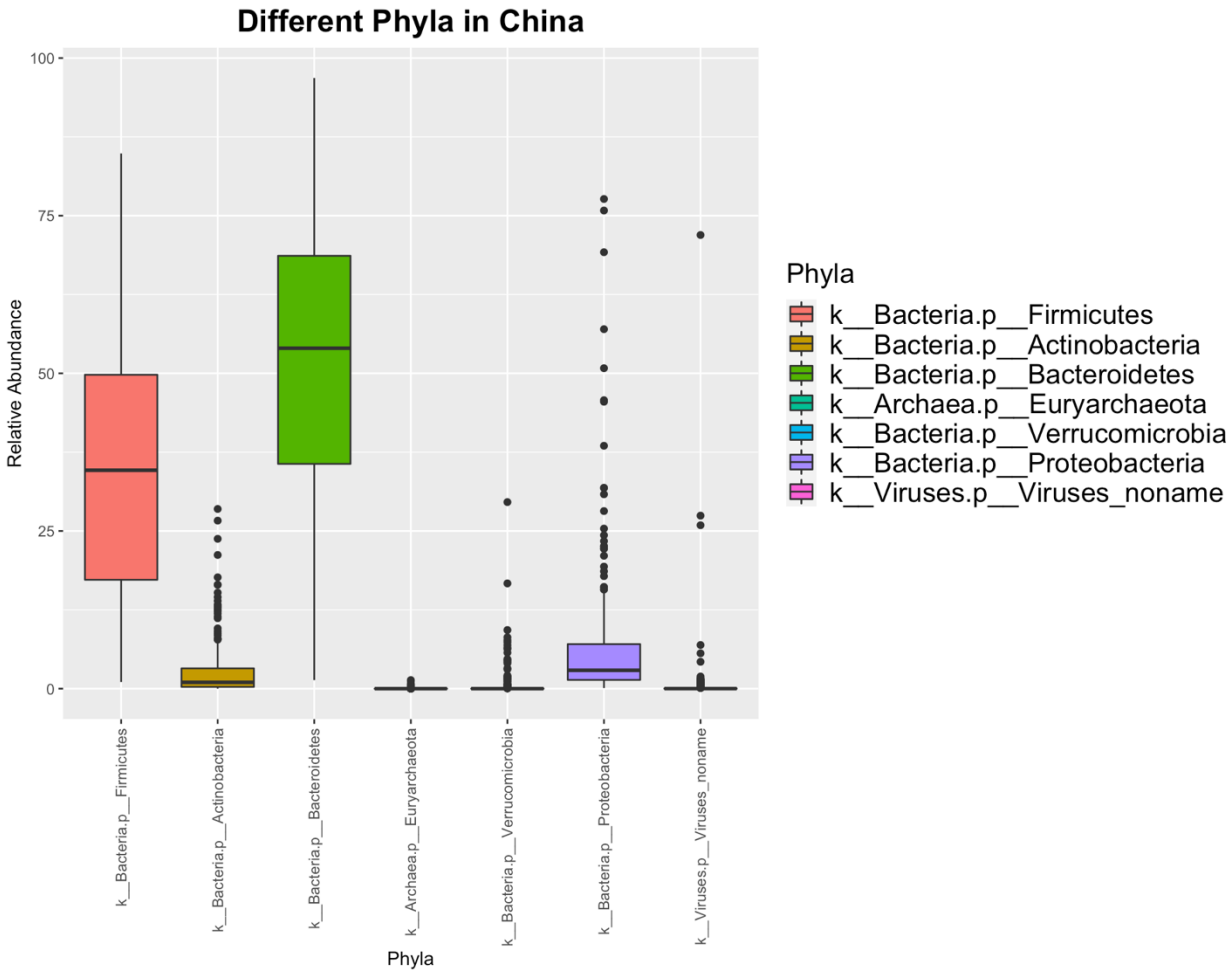

Figure 11: Phyla in healthy population from China.

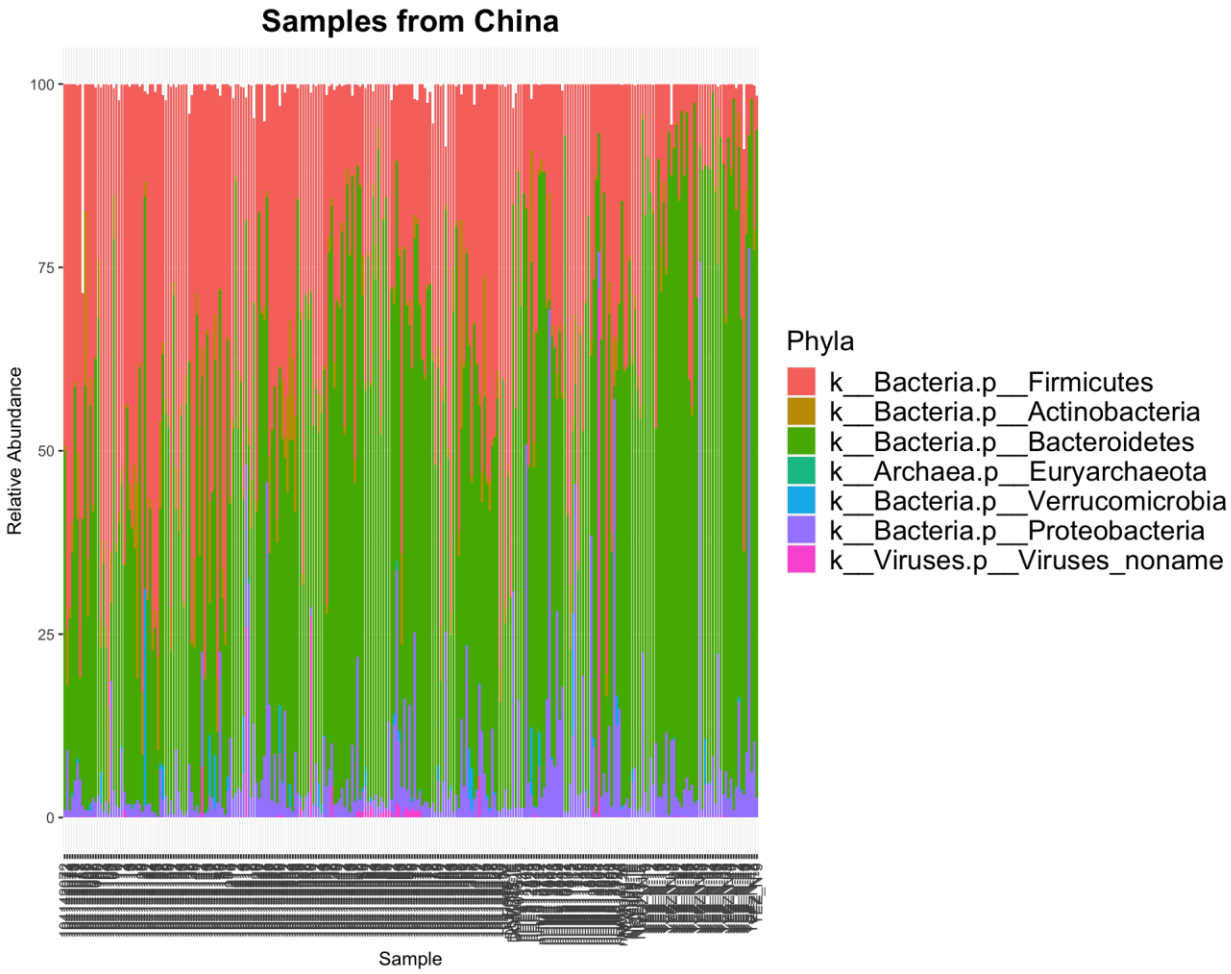

Figure 12: Relative Abundance of the different Phyla in China.

The analyzed data was for 268 healthy individuals from four studies (Jie et al., 2017; Jing Li et al., 2017; Junhua Li et al., 2014; Ye et al., 2018). Bacteroidetes had the highest relative abundance with a median of almost 55% (Figures 11 and 12). Bacteroidetes were followed by Firmicutes with a median of 30%. In contrast, Actinobacteria and Proteobacteria were rare, with Proteobacteria slightly more abundant than Actinobacteria.

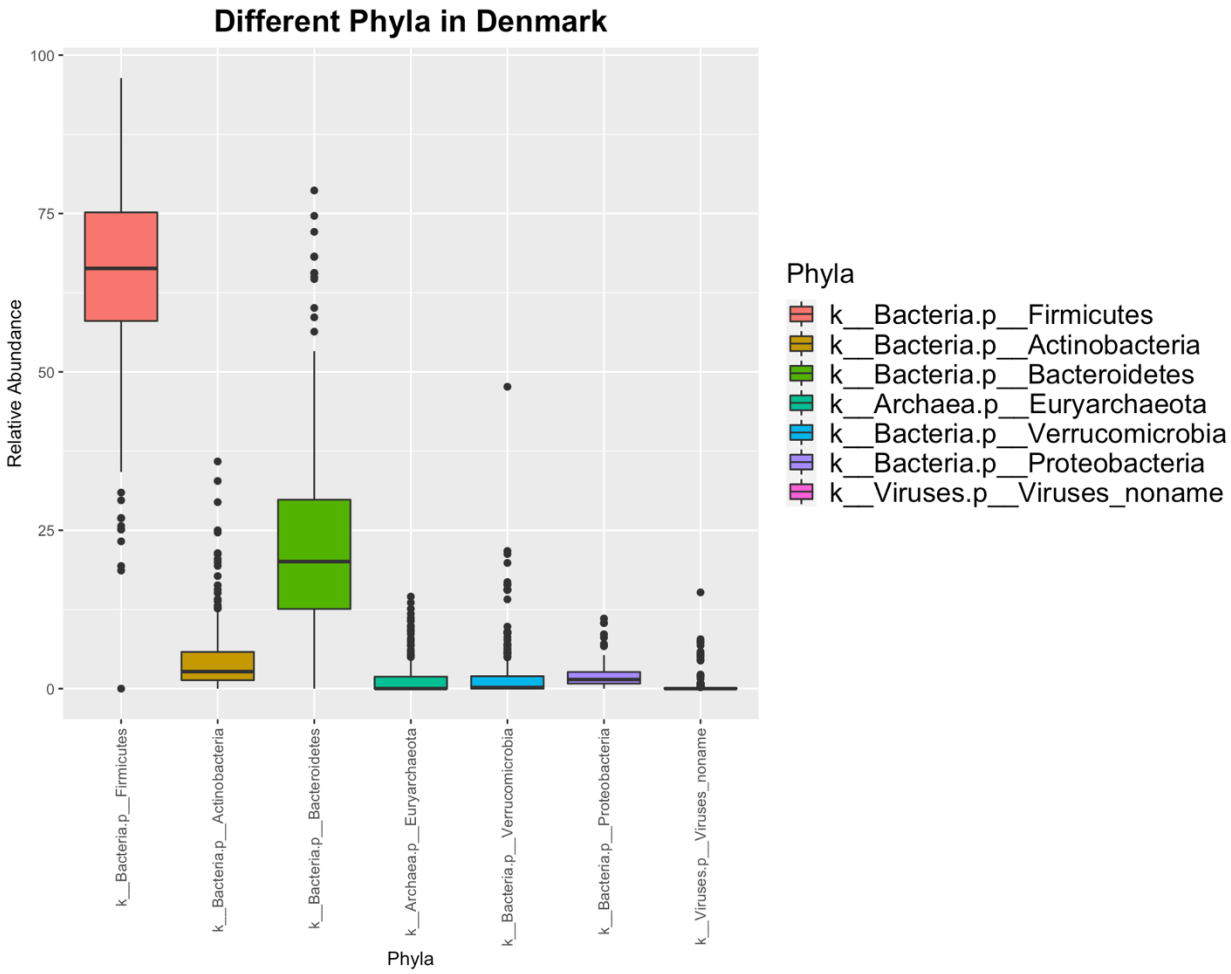

Figure 13: Phyla in healthy population from Denmark.

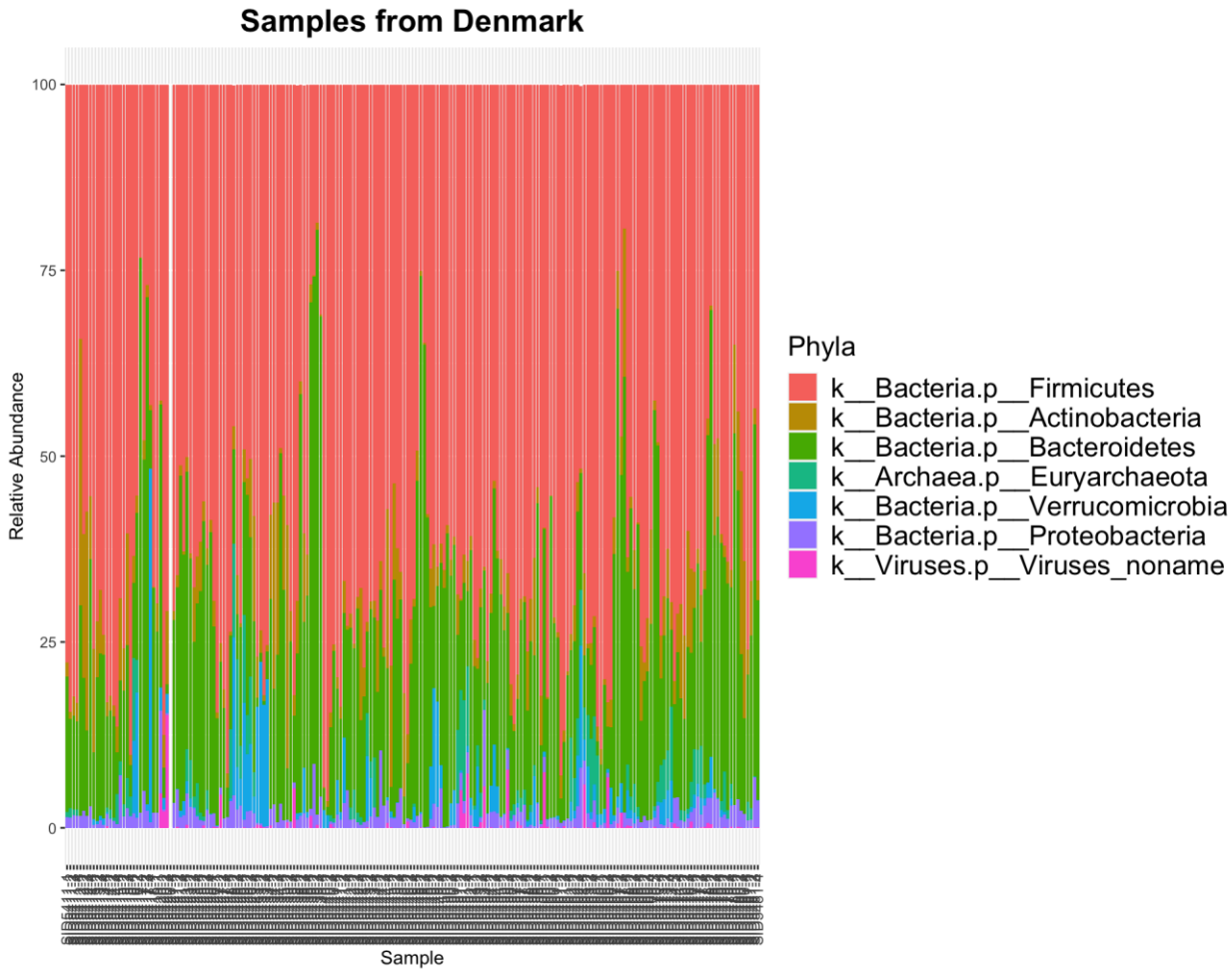

Figure 14: Relative Abundance of the different phyla in Denmark.

From the data of 208 healthy individuals in Denmark—116 females and 92 males—in the adult and senior age categories (Hansen et al., 2018), phylum Firmicutes had the highest relative abundance, with a median of 69% while Bacteroidetes had the second-highest relative abundance with a median of 20% (Figures 13 and 14). Other phyla had low abundance, in the descending order of: Actinobacteria, Proteobacteria, and Verrucomicrobia (Figures 13 and 14).

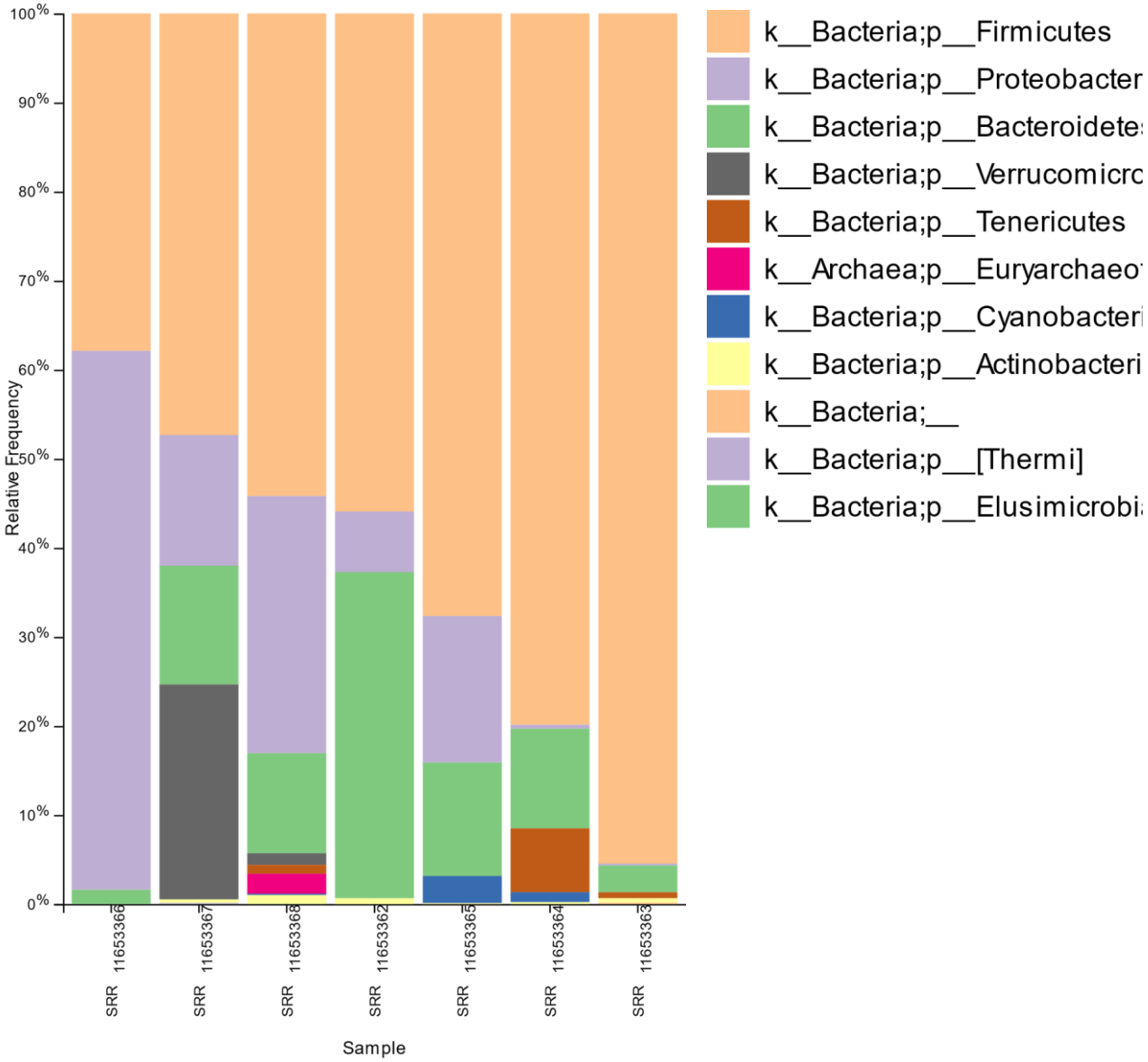

Figure 15: Absolute abundance Analysis of healthy microbiome from 7 Egyptian samples using QIIME2.

The percent abundance of different microbial phyla in seven stool samples from Egyptian adults (Radwan et al., 2020) was recomputed by QIIME 2 (Figures 13 and 14). Unlike all the past data sets, these data from Egypt was generated by 16S rRNA amplicon sequencing. It had high quality of 99.9%, and so using appropriate sampling depth resulted in keeping all the samples. The phylum of Firmicutes was the most abundant, followed by Proteobacteria and Bacteroidetes (Figures 14). The data was then combined to allow phylum comparison in all samples (Figures 16 and 17) to compare them to data from other countries.

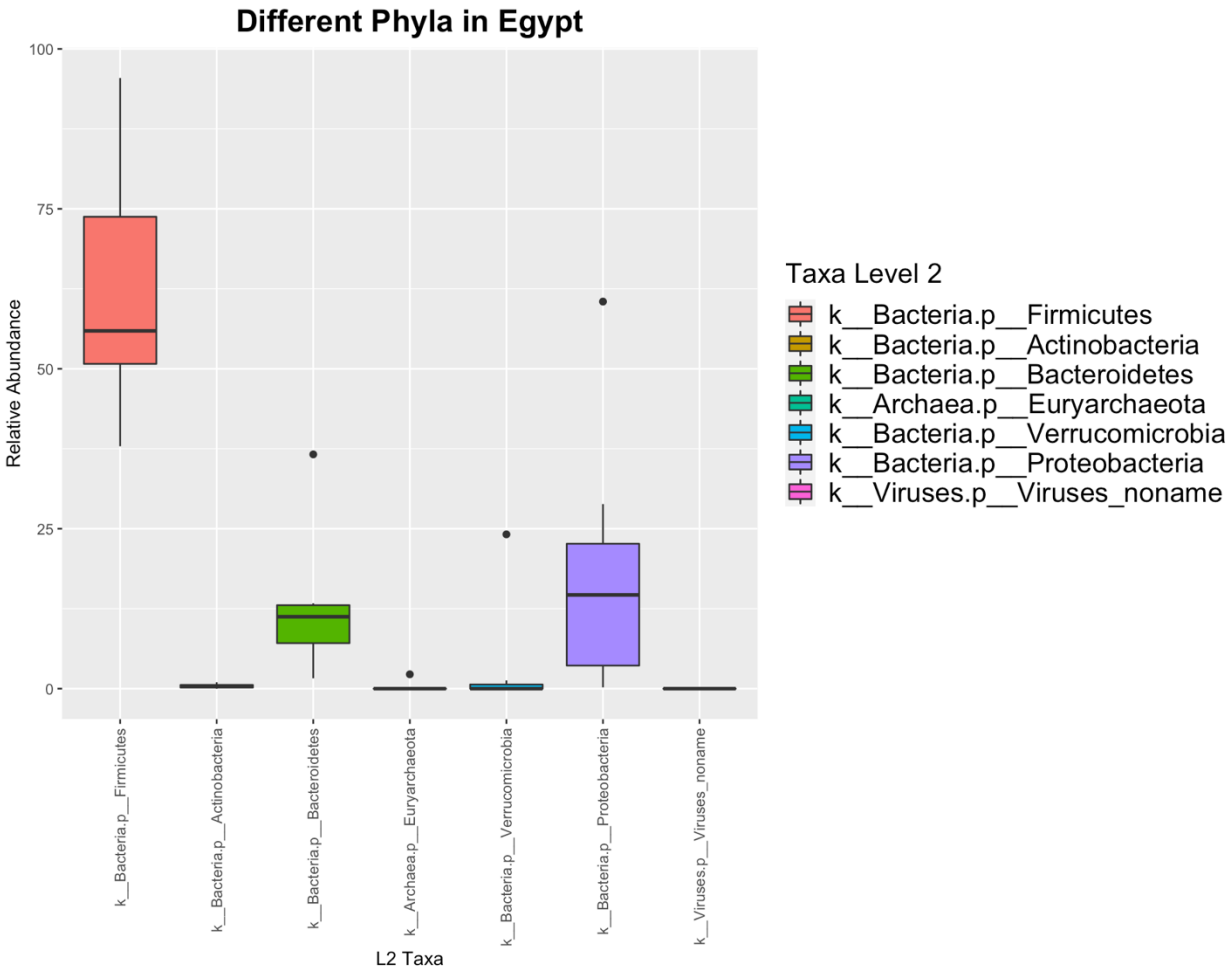

Figure 16: Phyla in healthy population from Egypt.

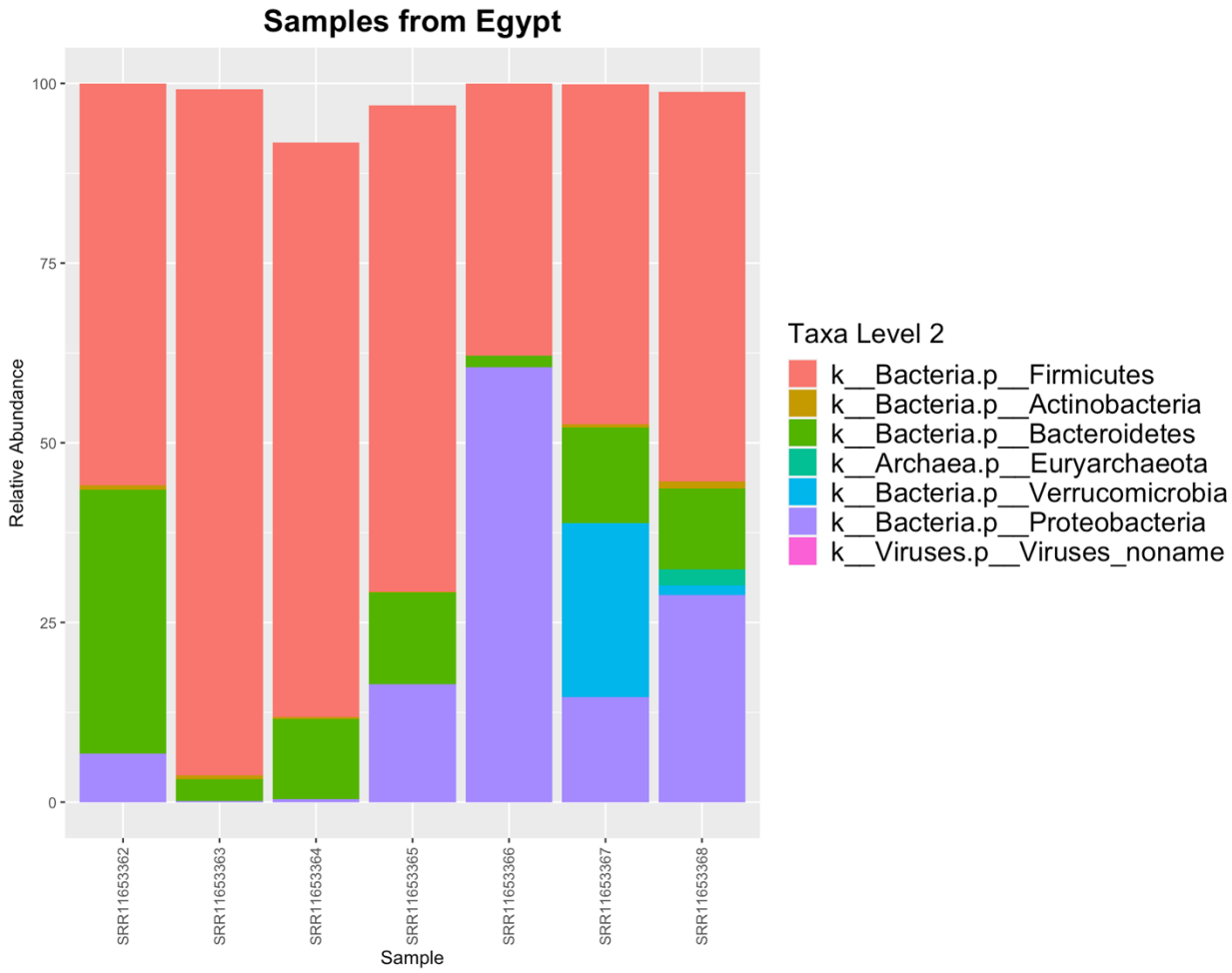

Figure 17: Relative Abundance of the different phyla in Egypt.

Firmicutes dominated the microbiome in the Egyptian samples with a median of about 55% (Figure 17). In addition, Proteobacteria was found to have an unusually high presence, with a median of almost 16%. Bacteroidetes had a much lower abundance with a median of nearly 10%. The other phyla were rare (almost 0%), including Actinobacteria (Figure 17). This may be odd compared to other countries, specifically for Actinobacteria, one of the four major gut phyla. Due to the low number of samples for Egypt and the reliance on data from only one study -despite being the only microbiome study from Egypt with data on SRA-, this may not be generalized over the whole Egyptian population.

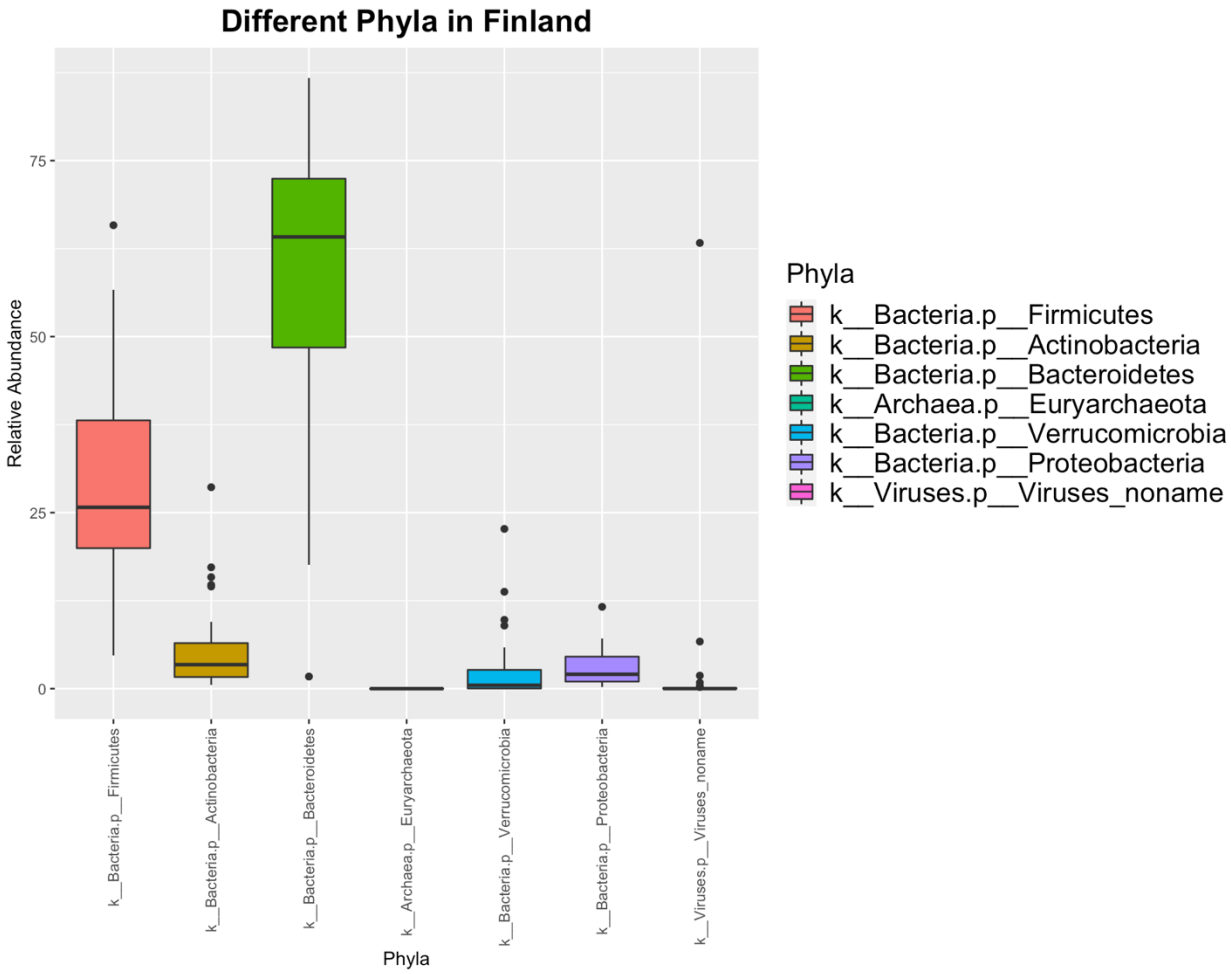

Figure 18: Phyla in healthy population from Finland.

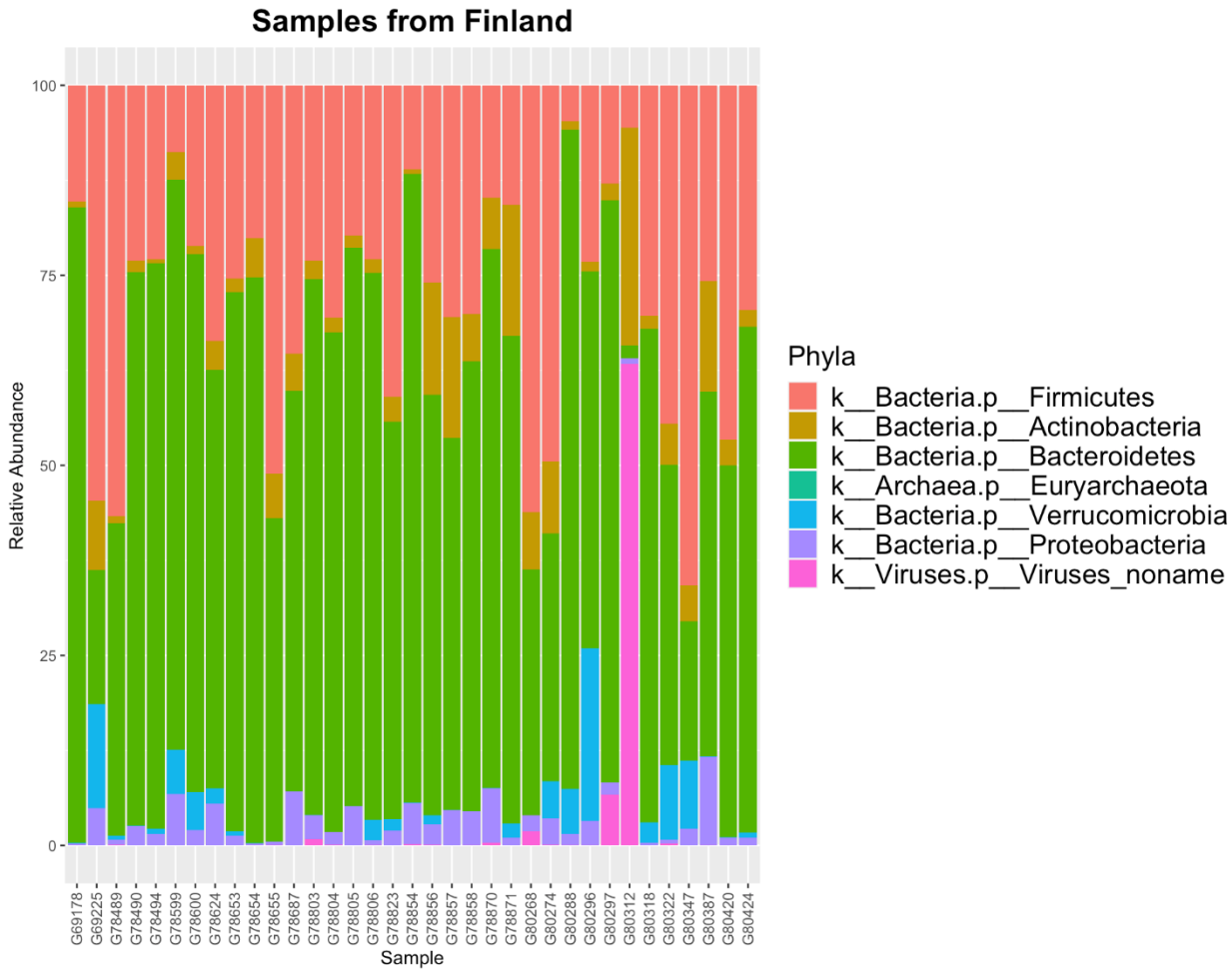

Figure 19: Phyla in healthy population from Finland.

The microbiome of 35 healthy individual children -13 females and 22 males- from Finland (Vatanen et al., 2016) was examined (figures 18 and 19). As shown in the figures, Bacteroidetes dominate the microbiome with a median of over 62%. Then Firmicutes came second with a median of 26%. The other phyla (Actinobacteria, Proteobacteria, and Verrucomicrobia) had very low relative abundance.

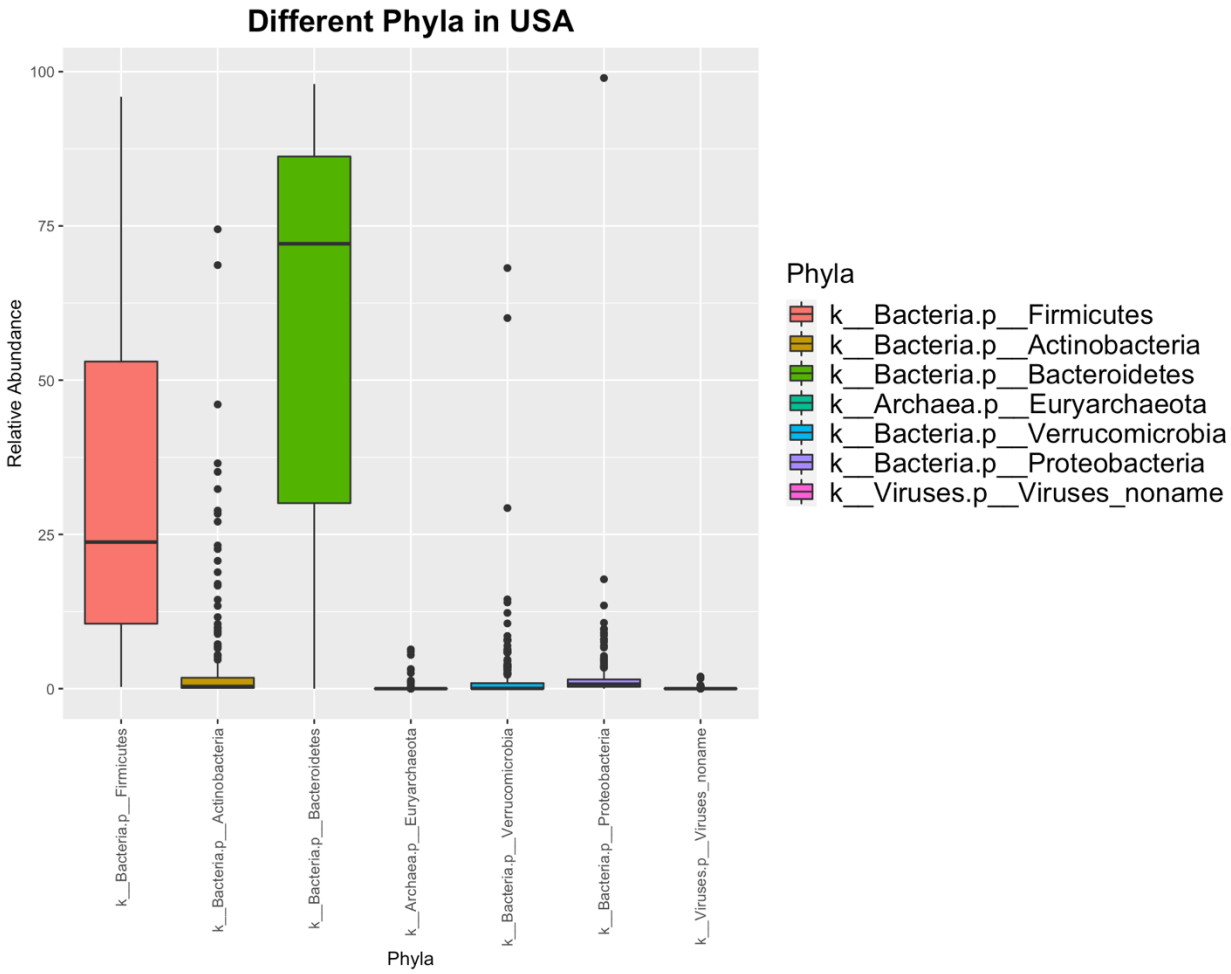

Figure 20: Phyla in healthy population from USA.

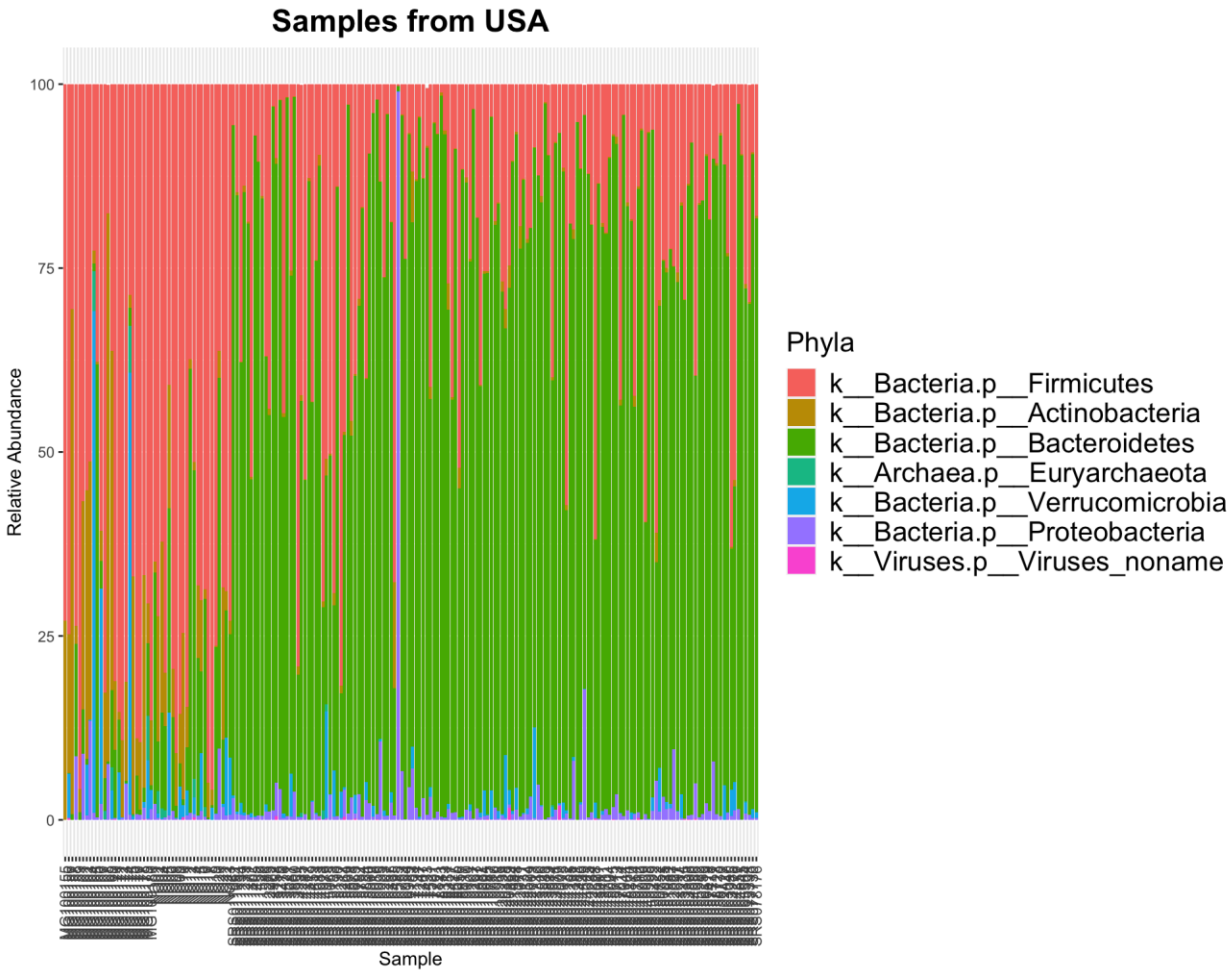

Figure 21: Phyla in healthy population from USA.

The microbiome of 194 healthy individuals (adults, child, and school-age) -males and females- from the USA (Hannigan et al., 2017; Huttenhower et al., 2012; Obregon-Tito et al., 2015) was examined. As shown in the figures (20 and 21), Bacteroidetes had the highest relative abundance in the samples with an extensive range that had a median of 73%. Then, Firmicutes had the second-highest relative abundance, spanned a vast field with a median of 24%. The other phyla, including Actinobacteria and Proteobacteria, had very poor abundance that was slightly above 0%.

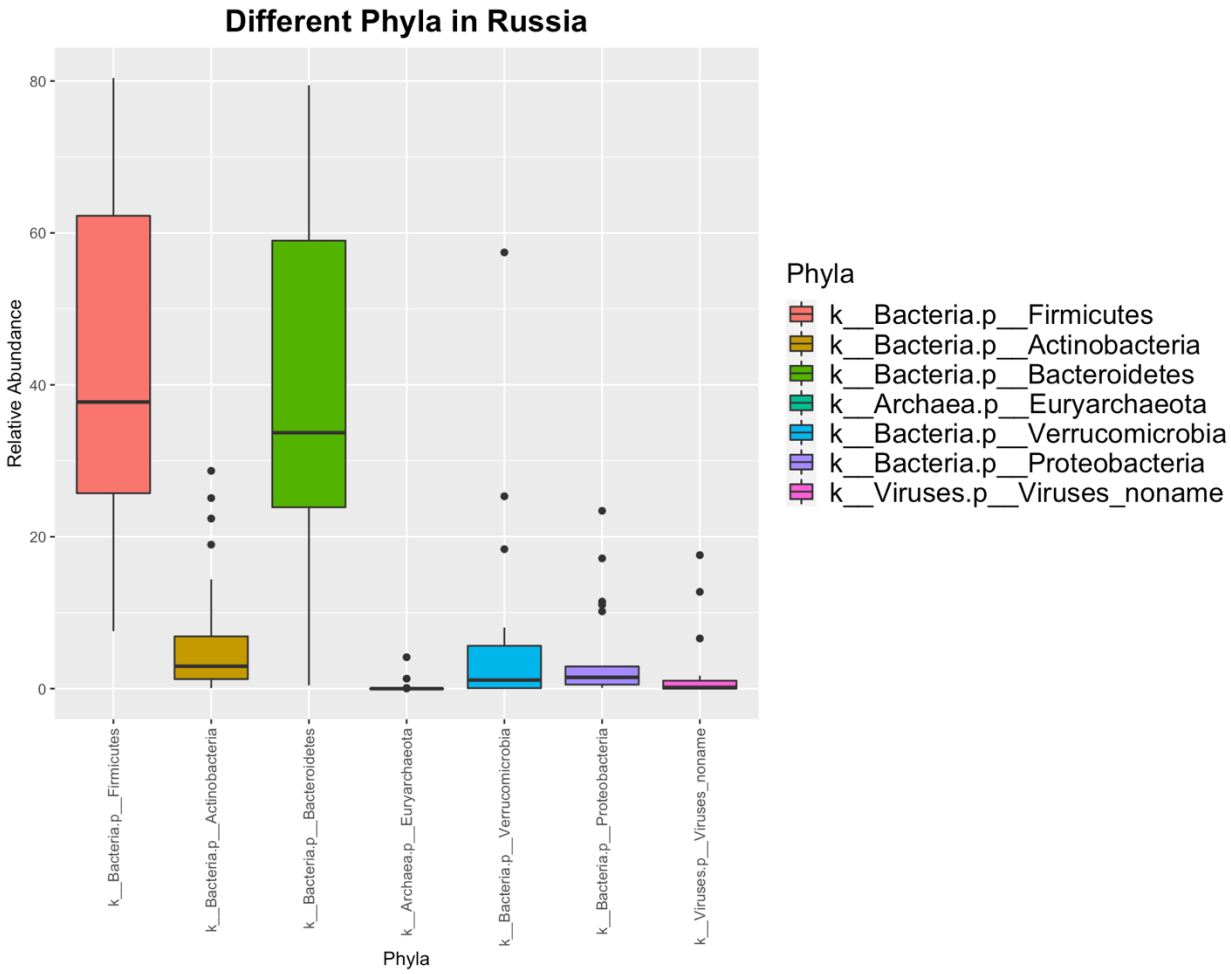

Figure 22: Phyla in healthy population from Russia.

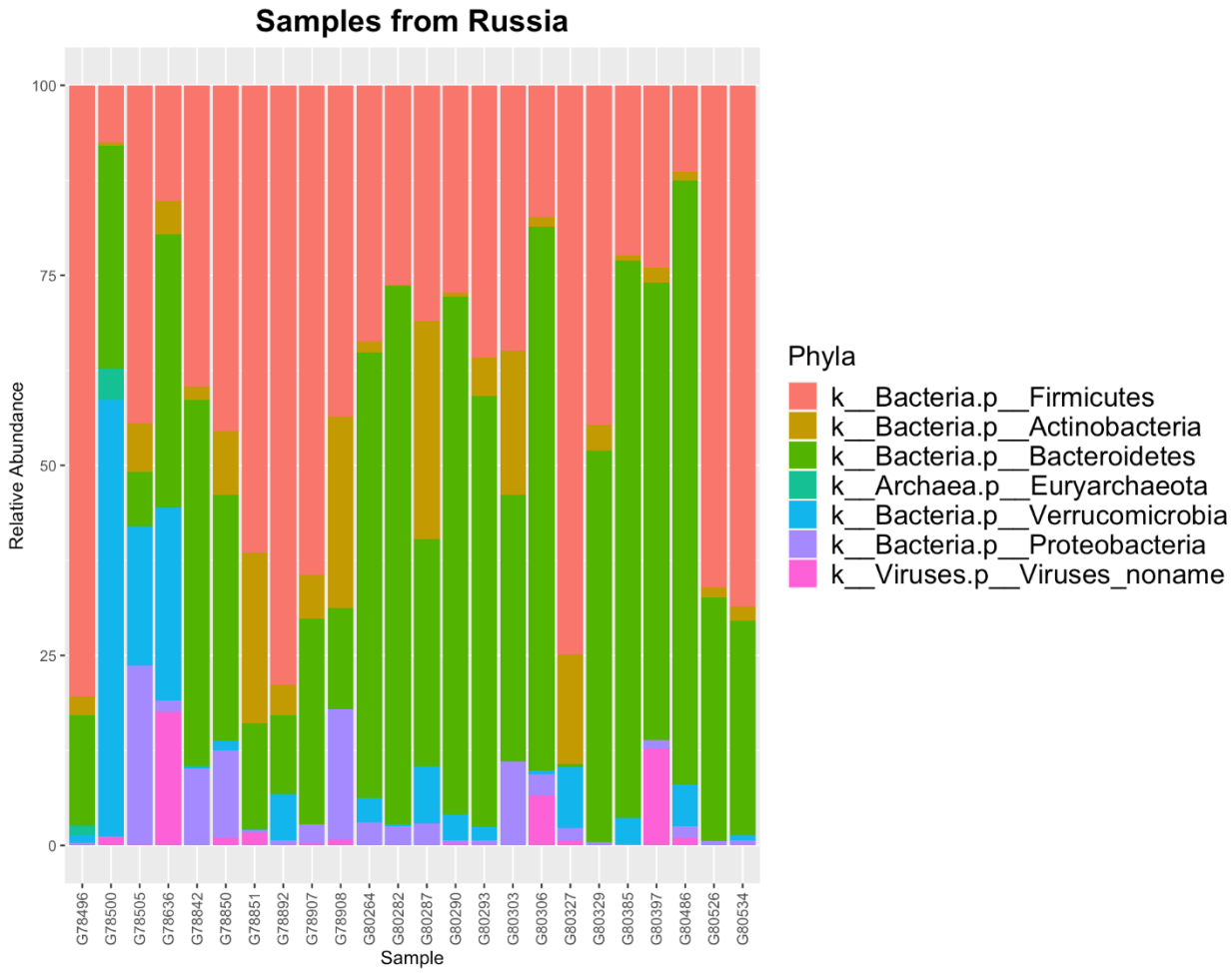

Figure 23: Phyla in healthy population from Russia.

The relative abundance data for 24 healthy individuals (Vatanen et al., 2016) was retrieved and used for the analysis as shown in figures (22) and (23). Both Firmicutes and Bacteroidetes had almost the same abundance with a very wide range with medians of 37% and 34%, respectively. In addition, both Actinobacteria, Verrocomicrobia, and Proteobacteria also had the same low relative abundance, with medians very close to 4%.

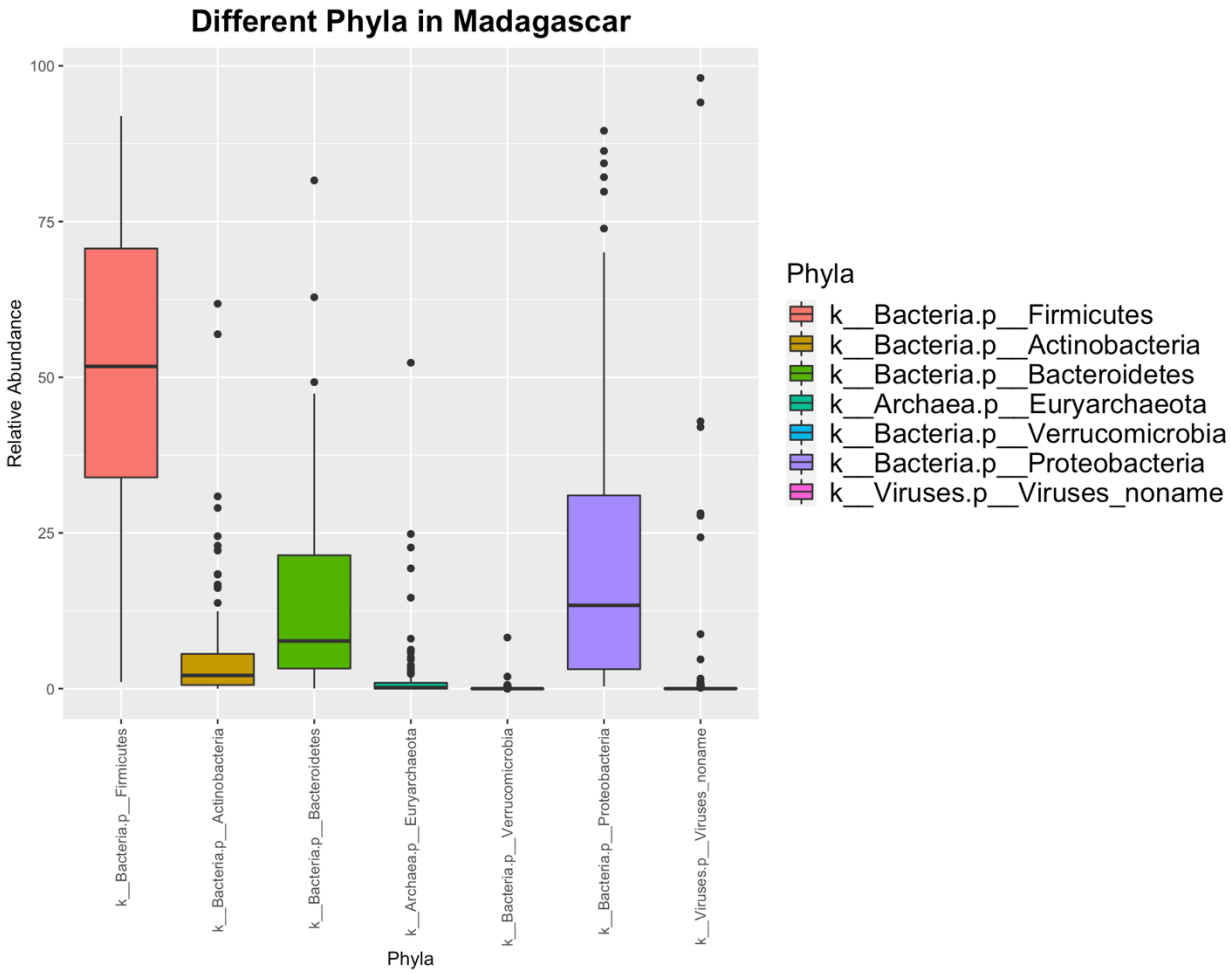

Figure 24: Phyla in healthy population from Madagascar.

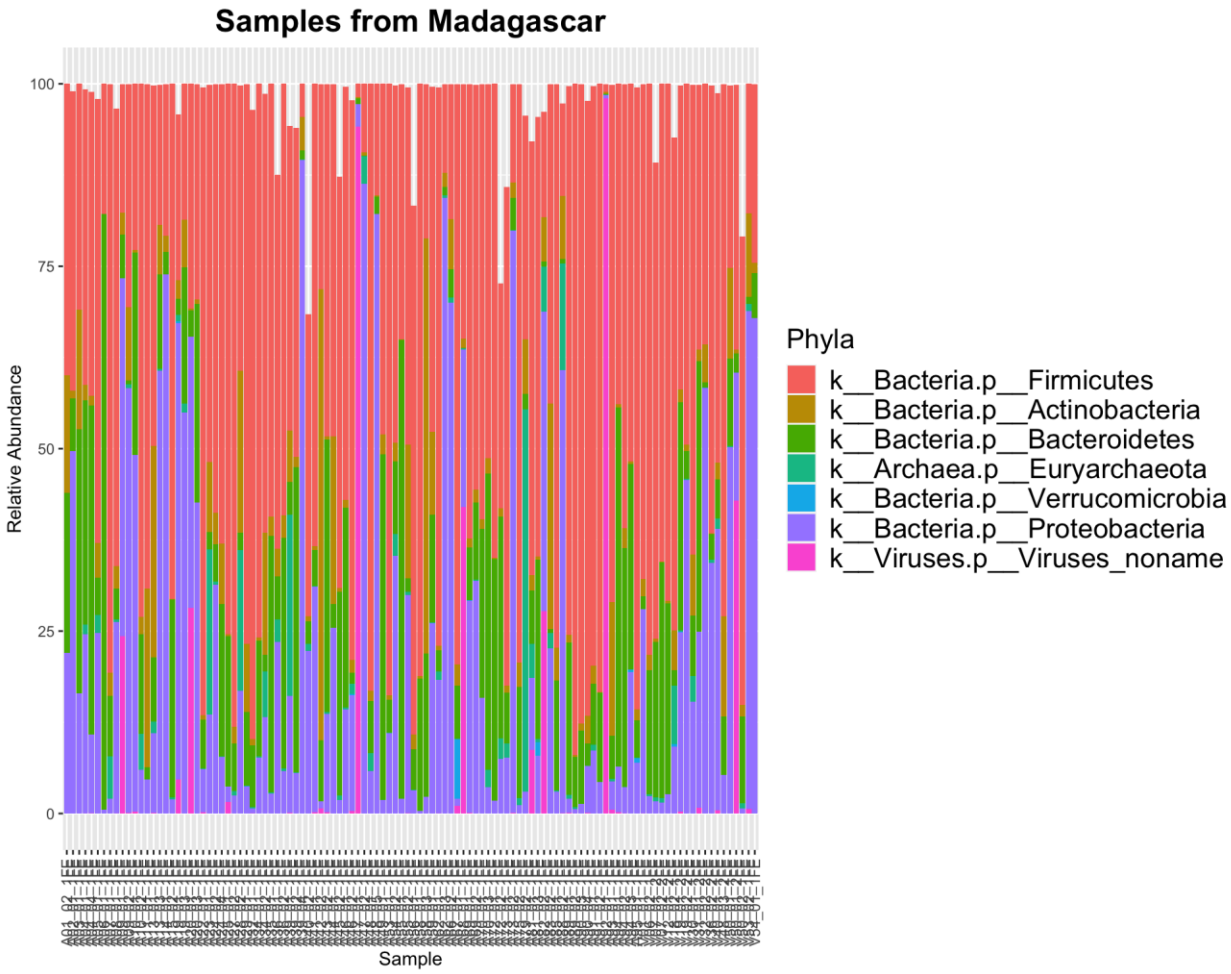

Figure 25: Phyla in healthy population from Madagascar.

The microbiomes of 112 healthy individuals from Madagascar (Seniors, Adults, and School-age)—males and females— (Pasolli et al., 2017) indicated that Firmicutes had the highest relative abundance, with a median of 52% (Figures 24 and 25). Proteobacteria was the second most abundant, with a median of 14%Bacteroidetes (median = 7%). Finally, Actinobacteria had a median relative abundance, close to 0%.

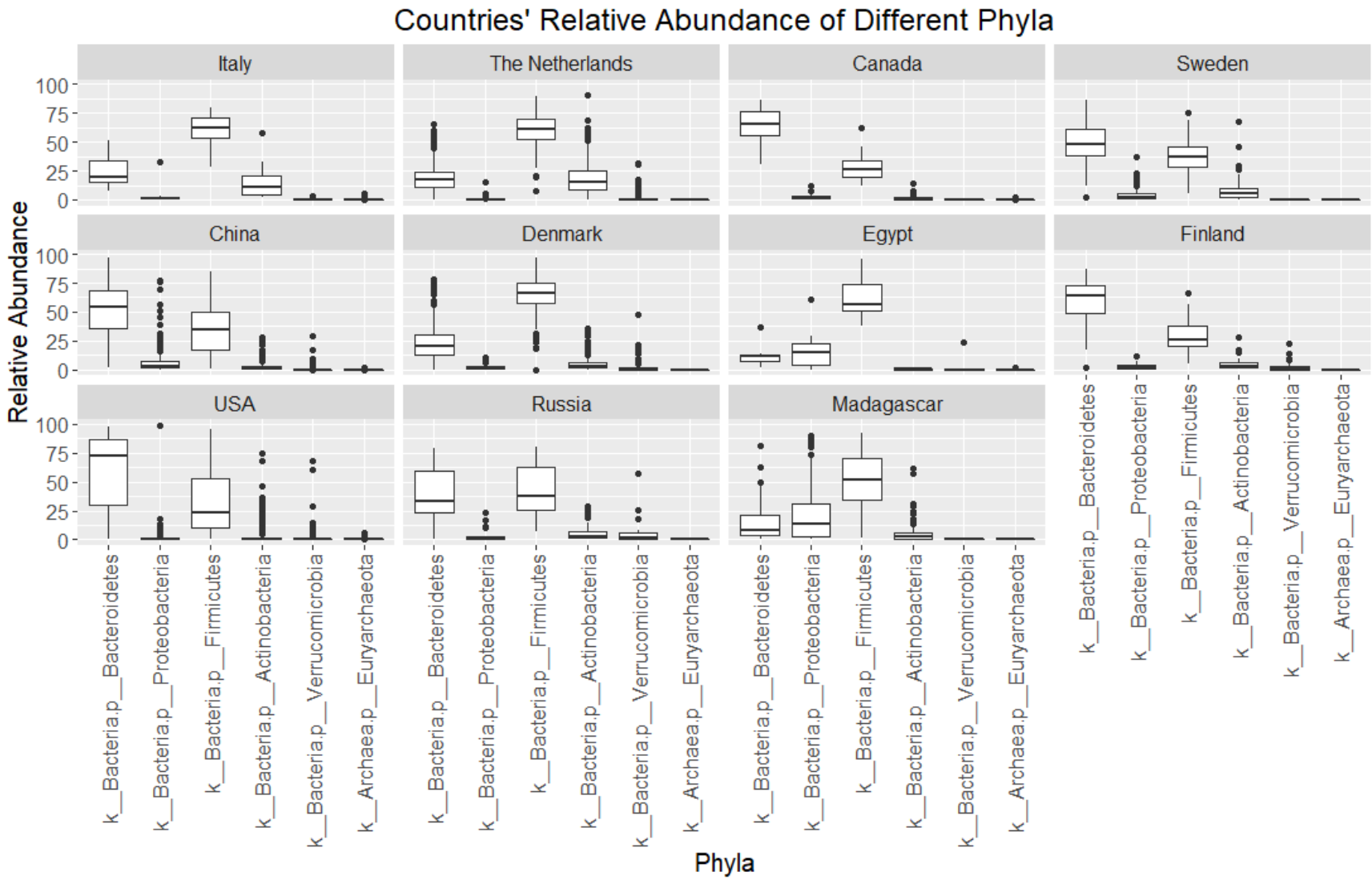

Figure 26: Prevalence of the different phyla in the samples from the countries in the study.

The prevalence of seven types of phyla was used to train the machine learning models (Figure 26). These were the phyla with the highest levels and were shared between the samples from the countries used in the study.

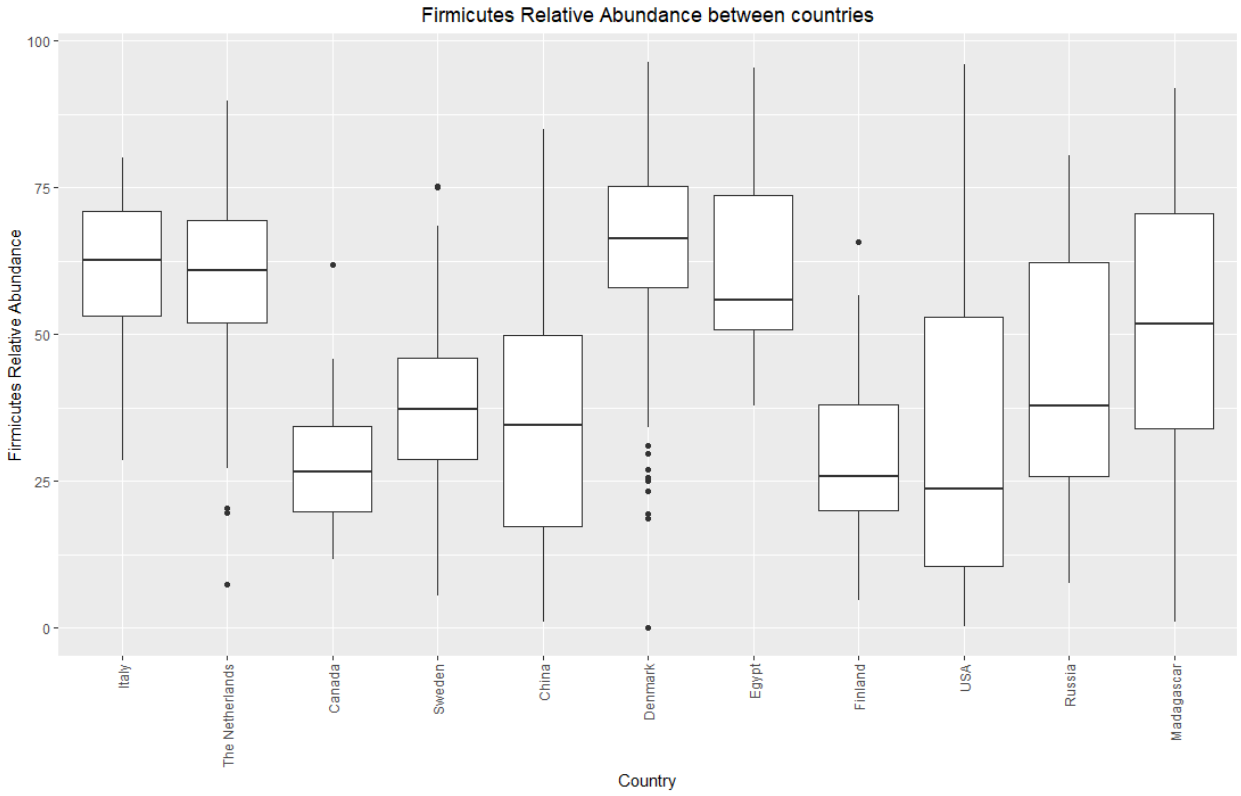

Figure 27: Comparison between the different countries used in the study for the prevalence of Firmicutes in their healthy microbiome.

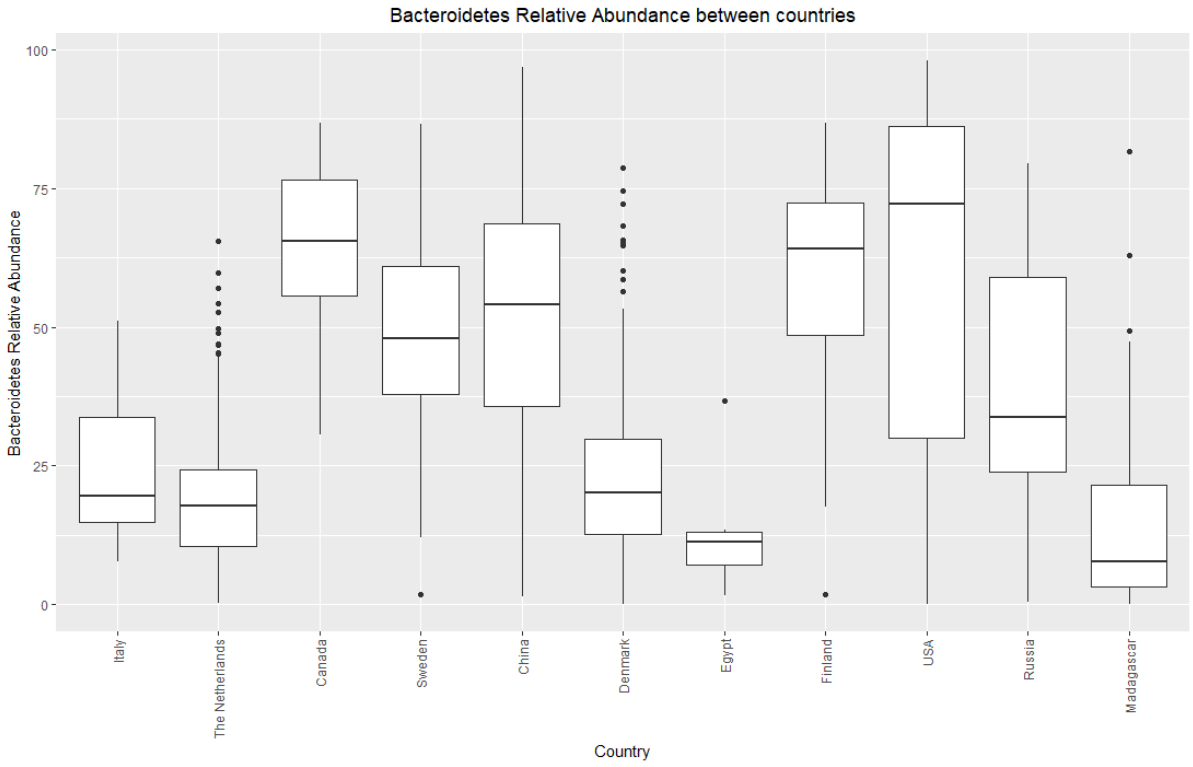

Figure 28: Comparison between the different countries used in the study for the prevalence of Bacteroidetes in their healthy microbiome.

Each of the major phyla was separately compared, for a better clarify (Figure 27) shows the relative abundance of the Firmicutes in the different countries used in the study. Firmicutes is one of the major phyla. As shown in the figure, it had a very high abundance in all countries. It was highest in Denmark followed by Egypt, Italy, Madagascar, and The Netherlands. It had lower abundance in Canada and Finland. Figure (28) shows the relative abundance of Bacteroidetes. It was highest in the USA followed by Canada. It was low in Egypt, Madagascar, and the Netherlands.

The gut microbiota plays a rule in the energy homeostasis through extracting the energy from food by fermentation and the formation of short chain fatty acids, SCFAs (Jumpertz et al., 2011; Turnbaugh et al., 2006). Based on many studies on animals and humans (Armougom et al., 2009; Bäckhed et al., 2004; Bervoets et al., 2013; Krajmalnik-Brown et al., 2012; Ley et al., 2005; Turnbaugh et al., 2009; P. Xu et al., 2012), it has been proposed that Firmicutes were more efficient in the extraction of energy from the food than Bacteroidetes. These data suggests that the changes in the bacterial composition/diversity are associated with the changes in the metabolic profile of the microbiota which affects the host's health. In the last decades, the Firmicutes/Bacteroidetes ratio was proposed as a hallmark for Overweight (De Bandt et al., 2011; Zou et al., 2020). Some studies (De Wit et al., 2012; Hildebrandt et al., 2009; Ley et al., 2006) proposed that the increase

of the abundance of Firmicutes over the abundance of Bacteroidetes leads to Overweight and when Bacteroidetes' abundance increases over Firmicutes there would be weight loss.

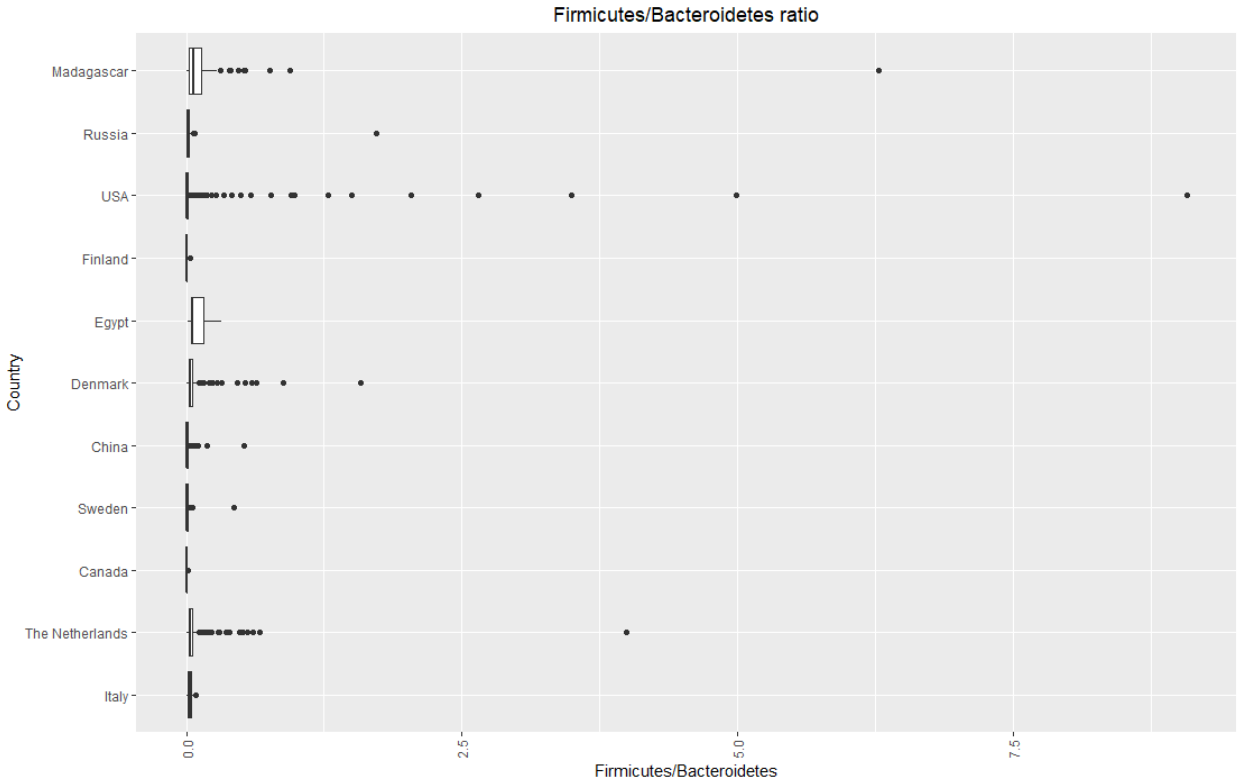

Figure 29: Firmicutes/Bacteroidetes ratio.

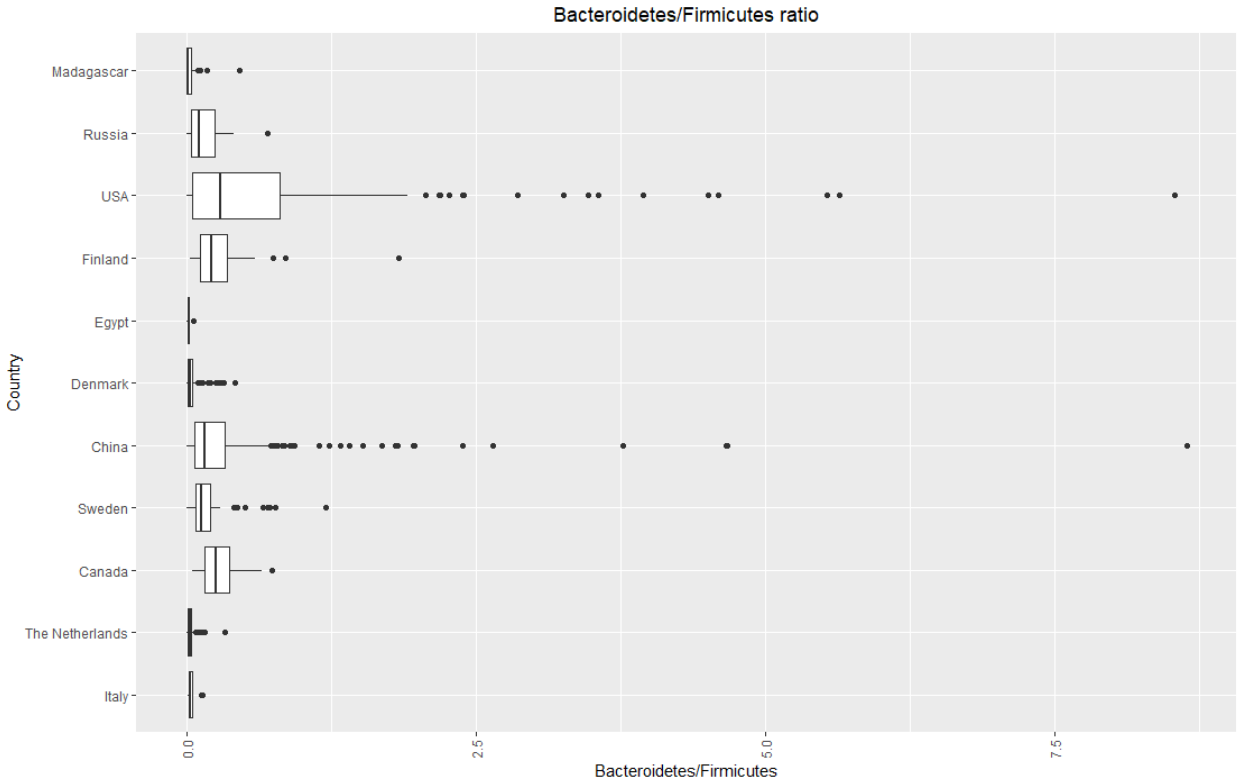

Figure 30: Bacteroidetes/Firmicutes ratio.

Bacteroidetes and Firmicutes produce different metabolites that have different effects on the host's body. Firmicutes produce butyrate while Bacteroidetes produce acetate and propionate (Fei & Zhao, 2013). In the colon, propionate stimulates GLP-1 and PYY release by L-entero-endocrine cells which leads to the inhibition of appetite (Chambers et al., 2015). It probably reaches the portal circulation, where it is captured by the liver and participates in gluconeogenesis and inhibit the expression of enzymes that are involved in the de novo synthesis of fatty acids and cholesterol (Demigné et al., 1995). Butyrate is a promoting the health (Den Besten et al., 2013) as it increases insulin sensitivity (Zhanguo Gao et al., 2009), exert anti-inflammatory activities (Säemann et al., 2000), regulate the metabolism in addition increasing the expression of leptin gene (Soliman et al., 2011). Acetate is absorbed and reaches the systemic circulation and the peripheral organs like the muscles, adipose tissue, and the brain. It works in contrast to propionate as it stimulates the synthesis of lipids (X. Gao et al., 2016). It activates the parasympathetic nervous system in the brain to promote the activation of insulin and ghrelin by the pancreas and the gastric mucosa, respectively (Perry et al., 2016). These events leads to the increased fat storage and appetite which contribute to Overweight (Magne et al., 2020). Relative abundance differences in Firmicutes and Bacteroidetes among countries (Figures 27 and 28) may indicate a possible correlation between the Firmicutes/Bacteroidetes ratio (Figures 29 and 30) and Overweight. Countries like the USA, Canada, and most of the European countries are known to have a large proportion of their population to be overweight, while many African countries like Madagascar are known to have a lower proportion of their population to be overweight (Sahned et al., 2019). Egypt can be excluded from this as the prevalence of overweight in Egyptian adults reached more

than 40% by 2019 (Aboulghate et al., 2021). This may be due to many factors like the type of diet and the health systems. The developed countries like the USA and The European countries depend high-protein high-fat diet in addition to better sanitation and hygiene practices (Greenhill et al., 2015; Mardanov et al., 2013; Sankaranarayanan et al., 2015; Tyakht et al., 2013).

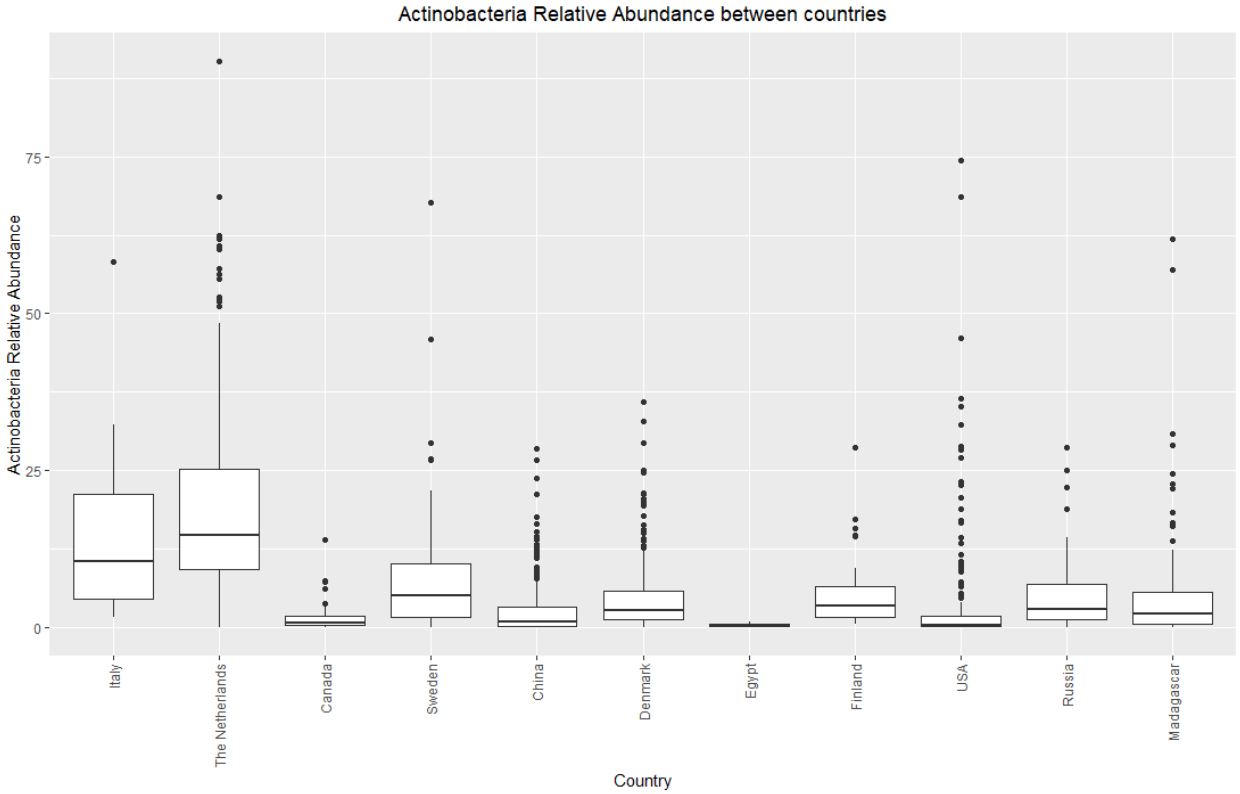

Figure 31: Comparison between the different countries used in the study for the prevalence of Actinobacteria in their healthy microbiome.

Actinobacteria was most abundant in the Netherlands followed by Italy (Figure 29). Despite being one of the major phyla, it was very low in Egypt, Canada, and the USA. The relationship between diet and Actinobacteria is still under investigation. Some studies demonstrated that the abundance of Actinobacteria is positively associated with a diet of high fats and negatively associated with the intake of fibers (Turnbaugh et al., 2008; G. D. Wu et al., 2011). On the contrary, other studies proposed an opposite correlation in which the increase abundance of Actinobacteria was positively correlated with being lean, the high consumption of complex carbohydrates, the improvements in glucose homeostasis, and reduction of obesity and inflammation (Cani, Neyrinck, et al., 2007; Geurts et al., 2014; Teixeira et al., 2013; Zimmer et al., 2012).

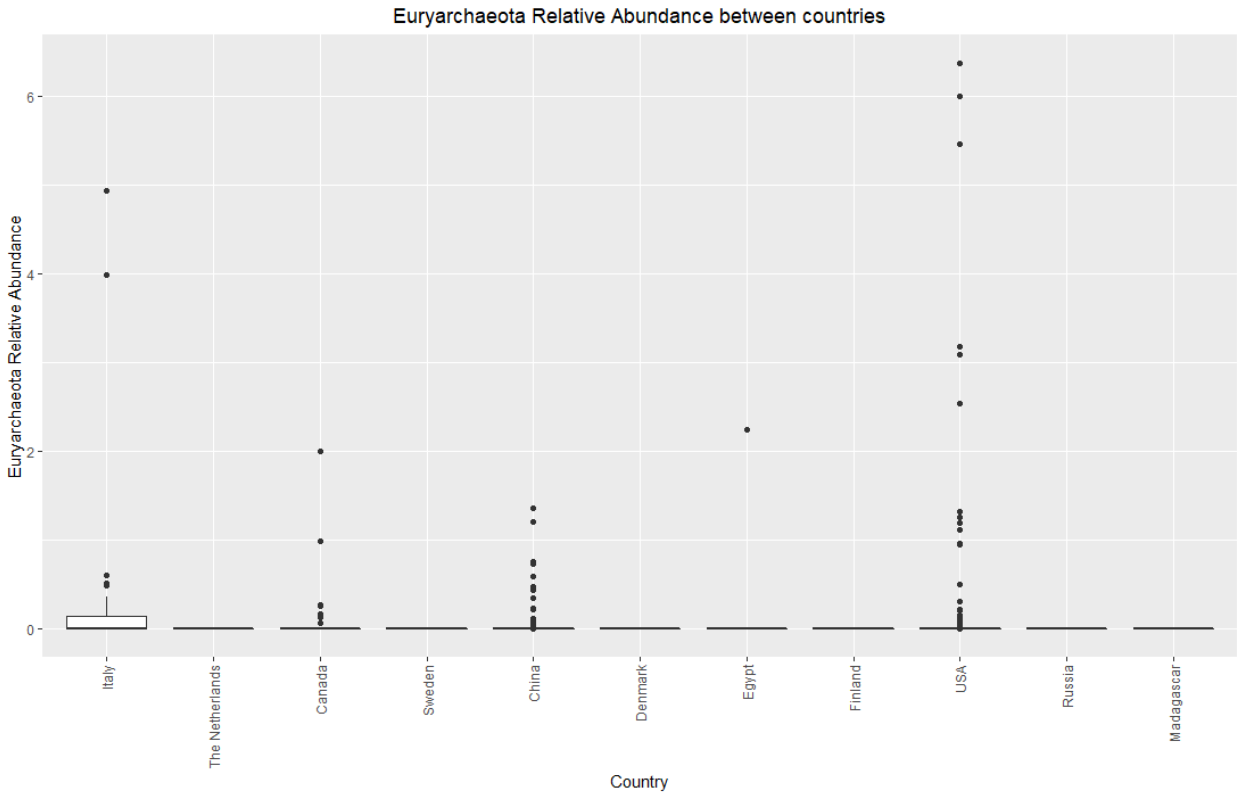

Figure 32: Comparison between the different countries used in the study for the prevalence of Euryarchaeota in their healthy microbiome.

Euryarchaeota was most abundant in Italy followed by the USA, and China (Figure 30). It was low in all the other countries. Archaea are known to constitute a low proportion of the human microbiome with low diversity. Only the phylum of Euryarchaeota is found to be part of the human microbiome with (*Methanobrevibacter smithii*, *M. oralis*, and *Methanosphaera stadtmanae*) as the main members. *Methanobrevibacter smithii* is the primary colonizer of the gut in humans (Horz & Conrads, 2010). A study proposed that despite being low in the human microbiome, archaea play a supporting role for the bacteria in the microbiome. The methanogens may aid in the process of interspecies hydrogen transfer to support fermenting bacteria in the gut which can be pathogens or at least opportunistic pathogens (de Macario & Macario, 2009). The data shown in the figure was from healthy subjects therefore this may explain the observed low abundance of archaea in all the samples.

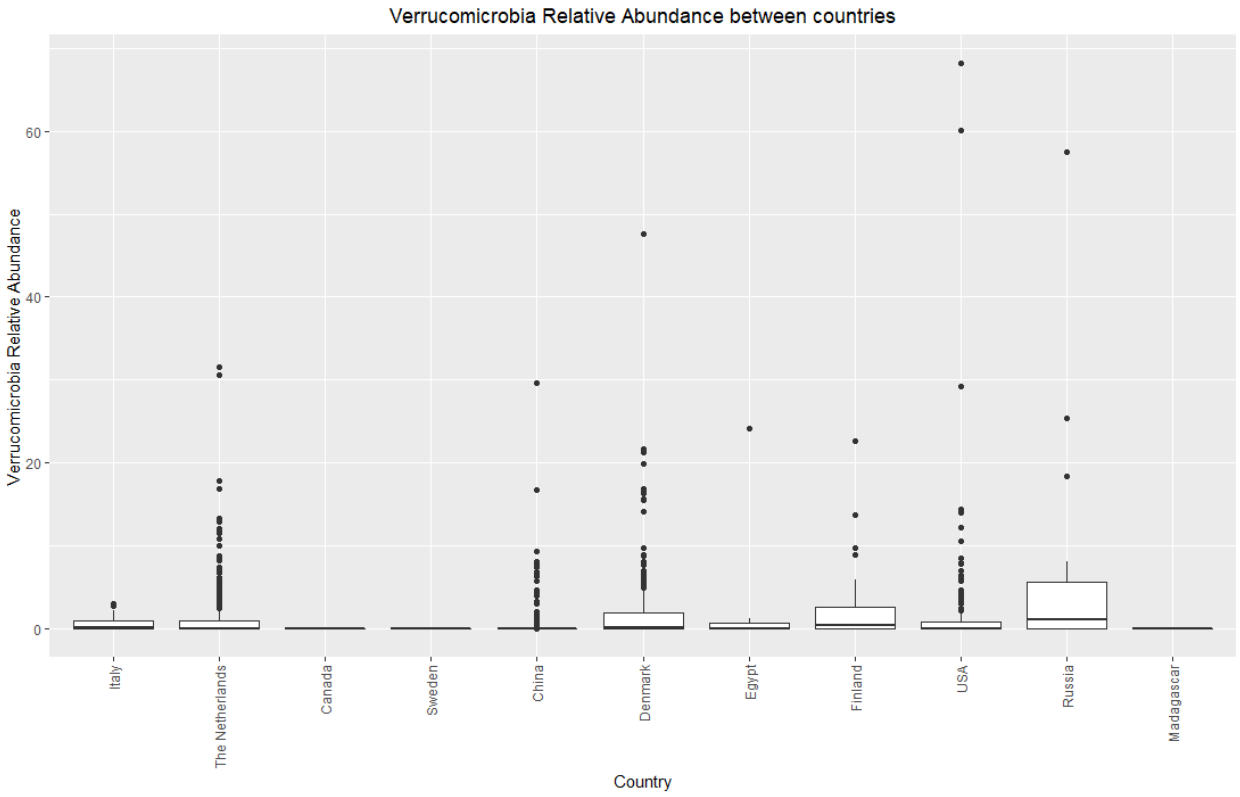

Figure 33: Comparison between the different countries used in the study for the prevalence of Verrucomicrobia in their healthy microbiome.

Verrucomicrobia was most abundant in Russia followed by Finland, Denmark, the Netherlands, Italy, and Egypt (Figure 31). It was lowest in Canada, Sweden, and Madagascar. Verrucomicrobia is a phylum with important members like *Akkermansia muciniphila* which is a mucin-degrading bacterium. It is believed play a role in the homeostasis of glucose and intestinal health (Belzer & De Vos, 2012; Johansson et al., 2011). It can represent 3% - 5% of the bacterial community (Santacruz et al., 2010). It resides mainly in the intestinal mucosa (Karlsson et al., 2012). According to many studies , its abundance is negatively correlated with body mass (Collado et al., 2008; Dao et al., 2016; Everard et al., 2011, 2013).

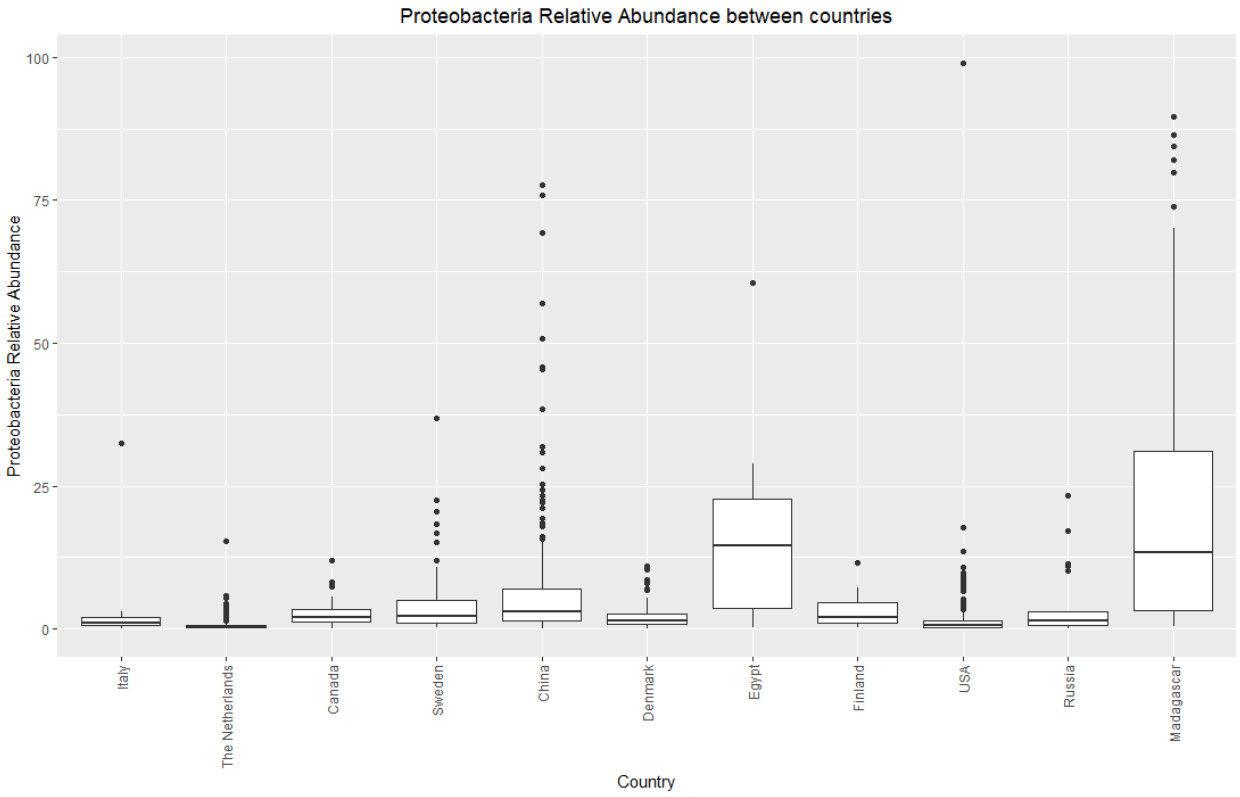

Figure 34: Comparison between the different countries used in the study for the prevalence of Proteobacteria in their healthy microbiome.

Proteobacteria had its highest relative abundance in Madagascar followed by Egypt, China, Sweden, Finland, Russia, Demark, and Canada (Figure 32). It was low in The Netherlands, The USA, and Italy. Although proteobacteria is one of the major phyla, it is associated with many human diseases. Proteobacteria is a phylum of Gram negative bacteria with lipopolysaccharide in the outer membrane (Rizzatti et al., 2017). There is an established correlation between low-grade inflammation sustained by lipopolysaccharides and the metabolic disorders (Hotamisligil, 2006). The production of lipopolysaccharides is sustained by gram negative bacteria in the gut - endotoxemia- and is reduced by the administration of antibiotics (Cani et al., 2008; Cani, Amar, et al., 2007). They have higher abundance mainly in countries from Africa and which have low quality health systems and higher consumption of antibiotics.

Machine learning models for geographical differences:

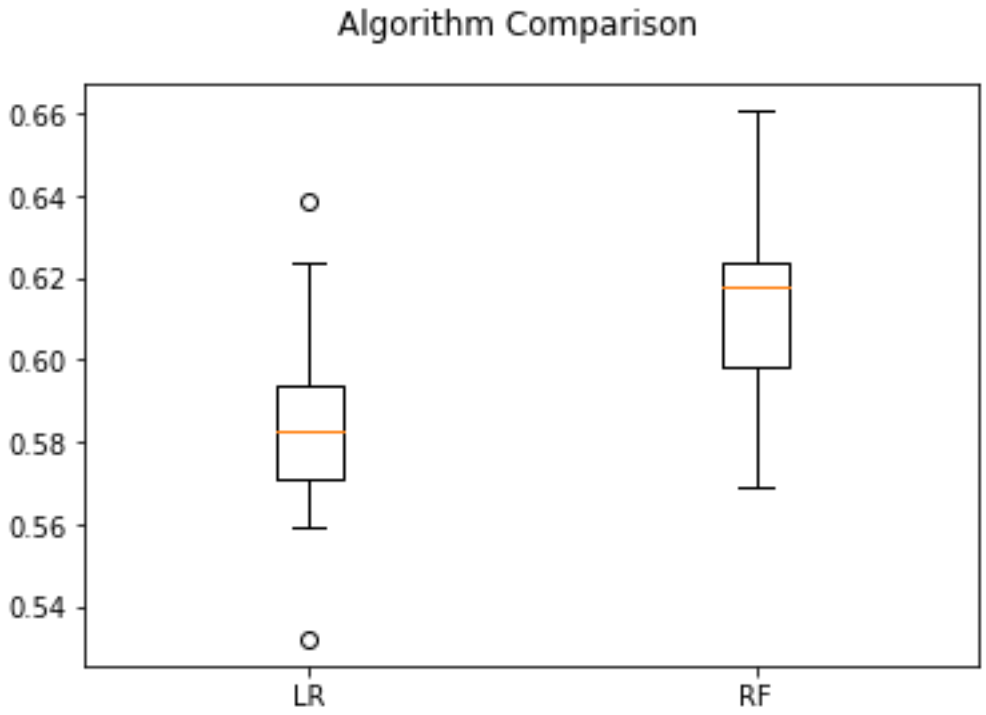

Figure 35: Comparison of the mean accuracy of Logistic Regression (LR) Classifier and random forests Classifier (RF).

Table 3: Random Forest vs Logistic Regression Classifiers.

	Logistic Regression Classifier	**Random Forests Classifier**
Mean Accuracy Score	58.499 %	61.4331 %
Standard Deviation	2.8901 %	2.3457 %

The random Forest classifier performed better than the Logistic Regression model, with a lower standard deviation (Table 3). Grid search was then used to optimize the Random Forest classifier's parameters. Random Forest classifier's performance was then improved from a mean accuracy score of 61.431% to a mean accuracy score of 65.84%.

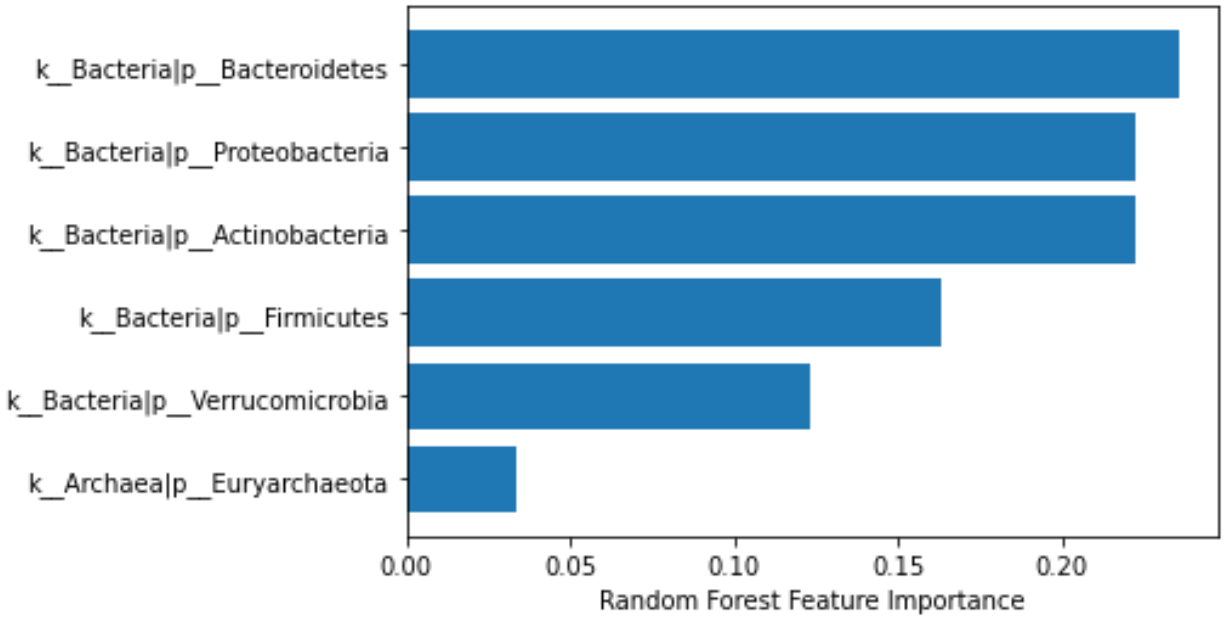

Figure 36: Using feature importance attribute in Random Forest classifier model method.

Table 4: Phylum vs Correlation Coefficient.

Phylum	**Correlation Coefficient**
Bacteroidetes	0.23594675
Proteobacteria	0.22206124
Actinobacteria	0.22203626
Firmicutes	0.16312876
Verrucomicrobia	0.12293449
Euryarchaeota	0.0338925

The most important features in the random forest classifier are Bacteroidetes, Proteobacteria, and Actinobacteria as the three most important phyla that affect the predictivity of the machine learning model, respectively. Phylum Firmicutes was fourth, followed by Verrucomicrobia, and Euryarchaeota, respectively.

COVID-19 related analysis

As the COVID-19 pandemic continues, the fight against SARS-CoV-2 will continue and even after the pandemic. The search for a correlation between the changes in the microbiome due to a specific disease is not a new objective. In this study, the aim was to test the changes that occur to the microbiome in patients with COVID-19 compared to healthy individuals from China.

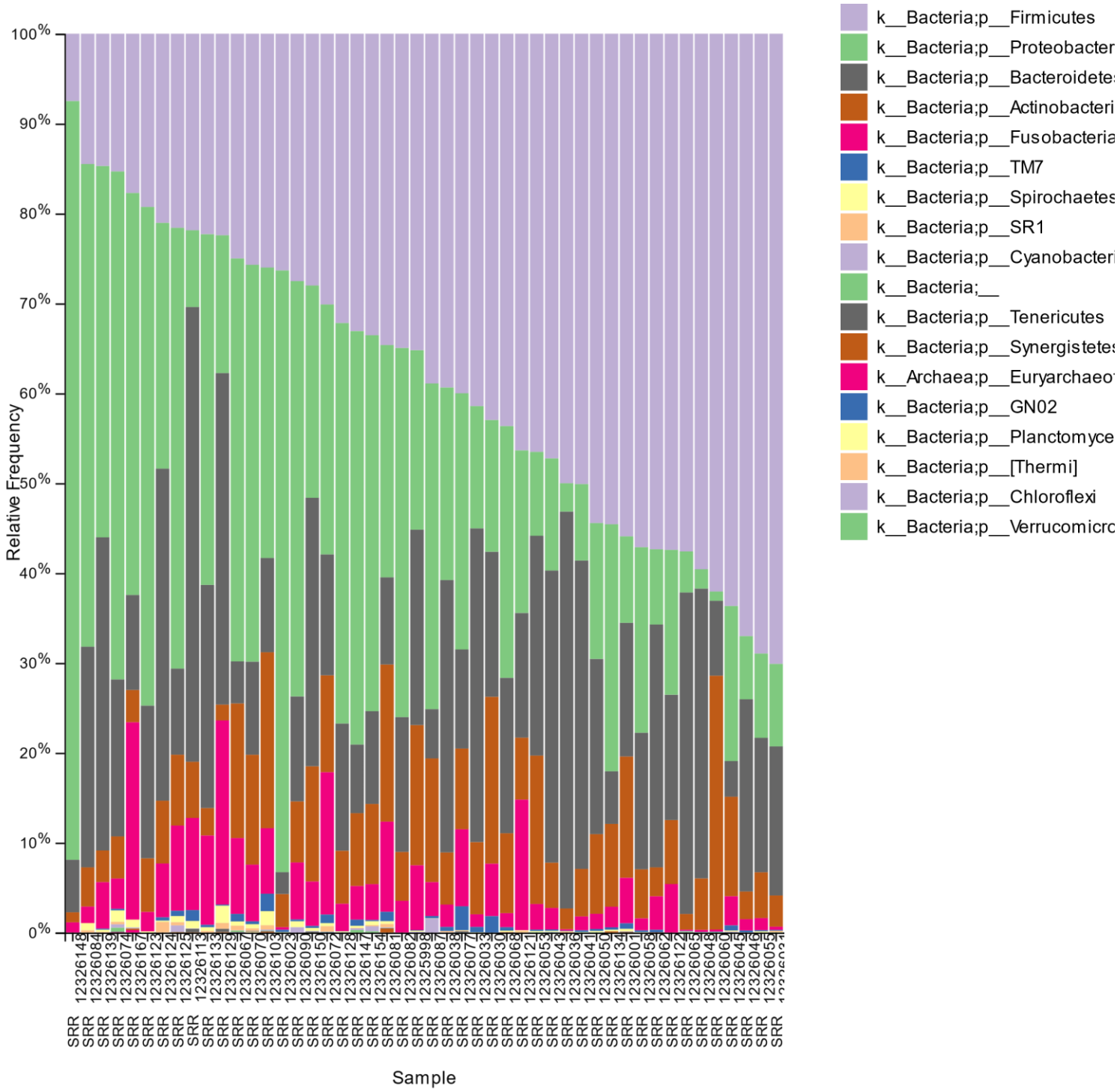

Figure 37: Absolute abundance of Phyla in Covid-19 population from China using QIIME2.

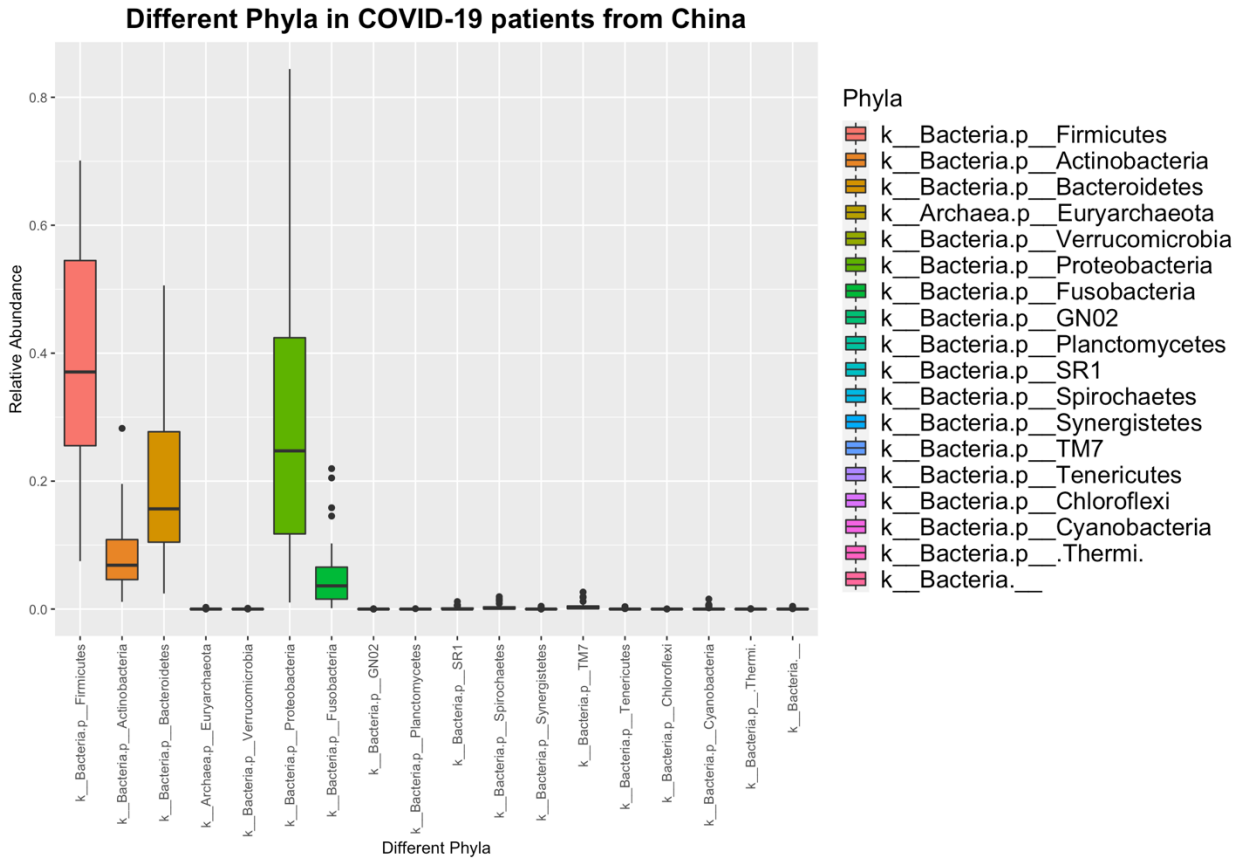

Figure 38: Phyla in Covid-19 population from China.

Figure (35) represents the absolute abundance on the phylum level for 47 COVID-19 patients from China. Figure (36) represents the relative abundance of these same phyla after transforming the absolute abundance into relative abundance. The microbiome of COVID-19 patients was analyzed in this study using the QIIME2 platform (Bolyen et al., 2019). Using the 16S rRNA amplicon sequencing data for stool samples from a study by Rong Xu and colleagues (R. Xu et al., 2020), The samples were isolated using anal swabs from 47 patients. As shown in figure (36), COVID-19 patients had Firmicutes as the bacteria with the highest relative abundance with a median of 37%, followed by Proteobacteria with almost 27%. Bacteroidetes came in third place with a relative abundance that has a median of 15%. Finally, Actinobacteria had the lowest relative abundance in the four major phyla, with a median of almost 9%. As shown in the figure, some other phyla had a feeble presence in the microbiome of the patients, with Fusobacteria having a higher presence than the other phyla except for the four major phyla.

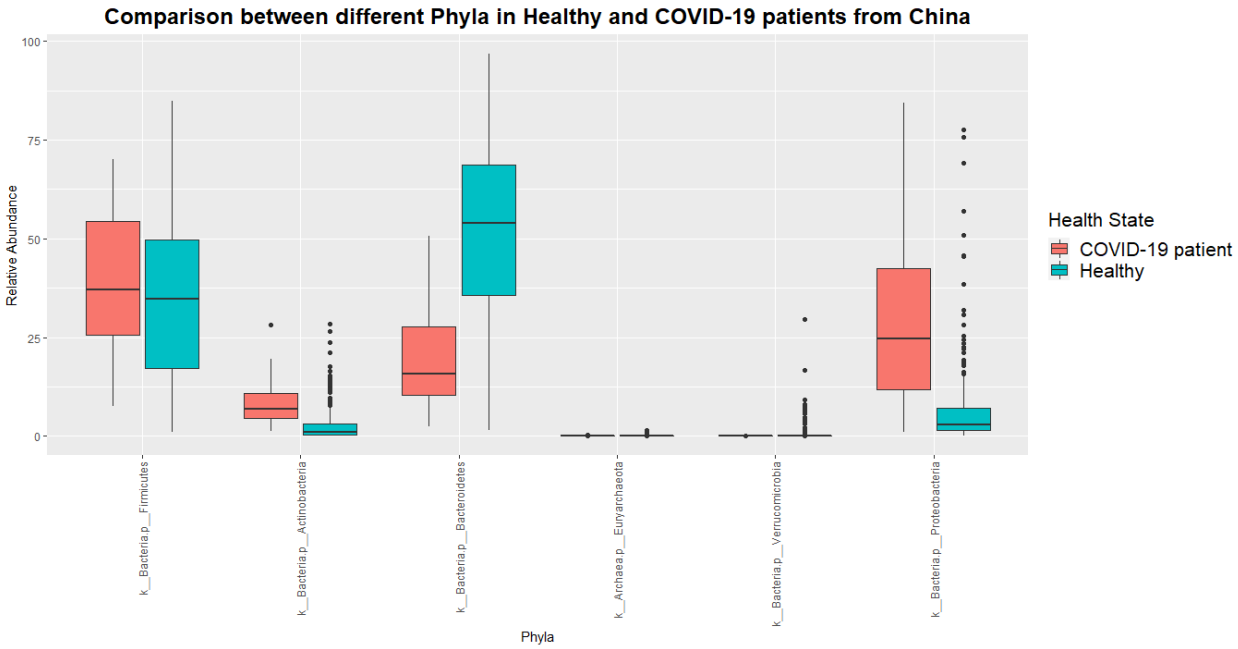

Figure 39: Comparison between shared phyla between healthy samples and Covid-19 patients from China.

Upon comparison with the data from the healthy individuals' studies figure (37), it was observed that Firmicutes was slightly higher in the COVID-19 patients than in healthy individuals. In addition, Actinobacteria had a higher abundance in the microbiome of the COVID-19 patients compared with healthy individuals. Also, Proteobacteria had a much higher abundance in the COVID-19 patients. On the contrary, Bacteroidetes had a much higher abundance in the healthy individuals with a median of 55% compared to 19% median in COVID-19 patients.

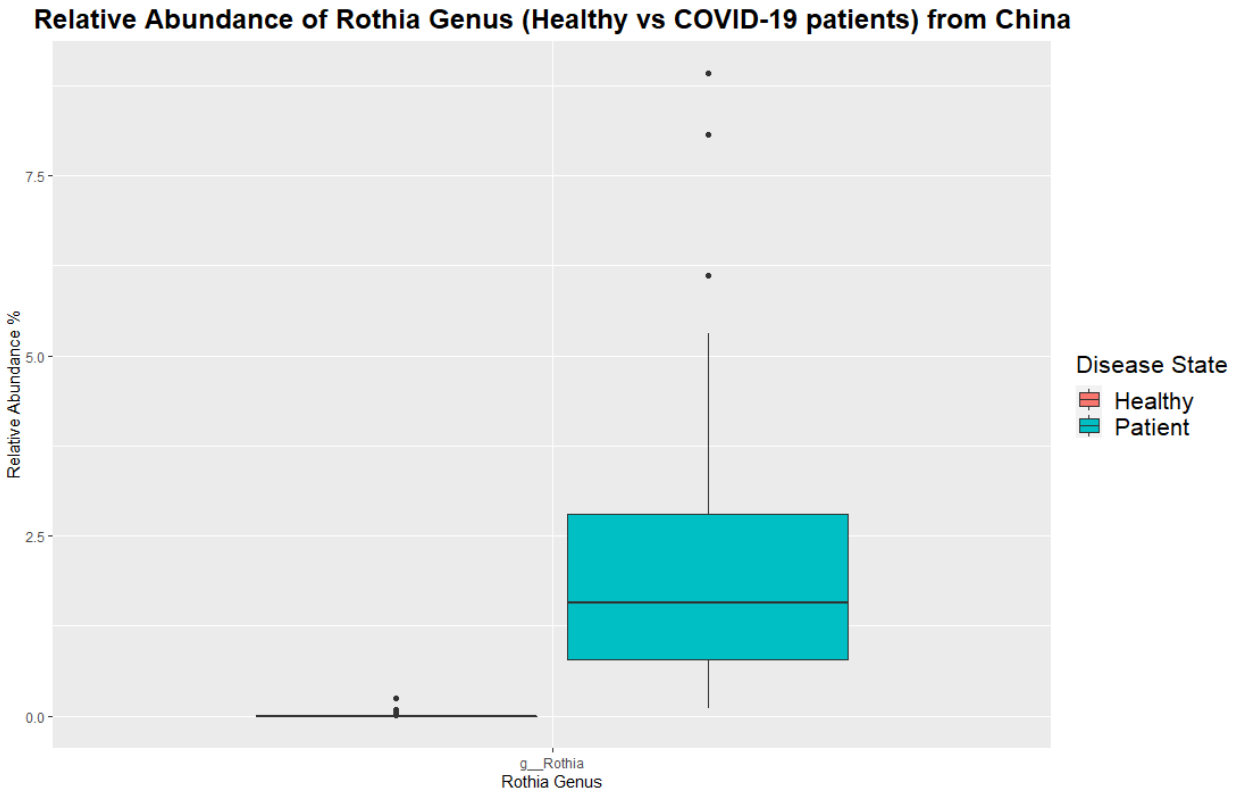

Figure 40: Relative Abundance of Rothia Genus in healthy samples vs COVID-19 samples from China.

The relative abundance of the Rothia genus in the healthy microbiome and patients with COVID-19 from China (Figure 38) suggests that the higher abundance of Actinobacteria in COVID-19 patients compared to healthy individuals can be due to the increase in the abundance of ***Rothia dentocariosa***, which is a species of Actinobacteria that as found to be associated with infection of SARS-CoV-2 by many studies (R. Xu et al., 2020; Zou et al., 2020). Rothia Genus is known to be in the oral microbiome of humans (Zaura et al., 2009). They were also classified as opportunistic pathogens (Boudewijns et al., 2003).

ACE2 is expressed abundantly in the intestines, specifically in the colonocytes of healthy individuals and patients with inflammatory bowel disease (J. Wang et al., 2020). It is able to regulate microbial ecology, amino acid transport, and gut inflammation (Hashimoto et al., 2012). Bacteroidetes were found to downregulate the expression of ACE2 in the colon of mice models. In addition, Firmicutes species had variable effects on the modulation of ACE2 expression (Geva-Zatorsky et al., 2017). A study on the microbiome of COVID-19 patients from Hong Kong (Zuo et al., 2020) found that some Bacteroidetes species had baseline abundance correlated negatively with the severity of COVID-19. In addition, Bacteroidetes had a higher abundance in the healthy samples from China compared with COVID-19 patients from China.

Proteobacteria were observed to have higher abundance in COVID-19 patients compared with Actinobacteria (Figure 38). This may indicate that the relative abundance of these phyla in healthy individuals and patients can be used as a marker for their infection with SARS-CoV-2. It needs further investigation as it may be a result for Dysbiosis and use of antibiotics. Actinobacteria is a phylum that can be associated with many health problems. It has some characteristics like fungi, as it can produce spores (Stackebrandt et al., 1997). Previous research has found that Actinobacteria can be very immunoactive, and so it can cause many respiratory disorders. For example, being exposed to high concentrations of some types of Actinomycete species can lead to allergic alveolitis (Falkinham III, 2003; Lacey & Crook, 1988; McNeil & Brown, 1994). In addition, Mycobacterium and Streptomyces species has been found to induce the production of proinflammatory cytokines and cause cytotoxicity in vitro and some systemic effects in vivo (Huttunen Maija-Riitta Hirvonen, Eila Iivanainen, Marja-Leena Katila, Kati, 2001; Huttunen et al., 2003; J Jussila et al., 2001, 2003; Juha Jussila et al., 2002). These studies would indicate a possible increase of Actinobacteria in COVID-19 patients, which was observed in the results of this study.

Proteobacteria phylum encompasses many human pathogens like Rickettsia and Brucella, both genera from the Alphaproteobacterial class, Neisseria and Bordetella, which belong to Betaproteobacteria class in addition to others in Gammaproteobacteria class and Epsilobacteria class like Escherichia and Helicobacter, respectively. Proteobacteria are found in many places in the human body, such as skin, tongue, oral cavity, and vaginal tract, in addition to the human gut and stool (Huttenhower et al., 2012). When comparing their abundance in the COVID-19 patients, it had higher abundance as mentioned above, which can be explained due to the alterations that occur in the microbiome of COVID-19 patients due to the infection. This point is in agreement with different studies that detected changes in the microbiome of COVID-19 patients compared to healthy samples (R. Xu et al., 2020; Zuo et al., 2020).

Chapter 4: Conclusions and Future Perspectives

Both machine learning models had high predictivity, which indicates a correlation between the differences in healthy microbiomes and their population of origin. It may also indicate batch and age effects.

The microbiome predisposes individuals to infection to different diseases. There are many factors contributing to the differences in the population-based microbiome variations. These factors are like diet, exposure to pathogens, age, psychological stress and anxiety, smoking and alcohol consumption. On the country level, these factors can be summarized as the differences in their lifestyles and their health systems. The study of the phyla can indicate new markers for the microbiome changes due to SARS-CoV-2. As observed in this study, the trend of Actinobacteria, Proteobacteria and Bacteroidetes in healthy individuals was almost flipped in the COVID-19 patients. This relationship needs further investigation.

The study may have some limitations that require further investigations in the future. The data should have been for the same age categories. More countries should be added to widen the training and test sets so the machine learning models would have better predictivity. In addition, more samples for countries like Egypt would have made the analysis more accurate and will reduce the batch effect.

In addition, working on more profound microbiome levels, like working on Class, Order, Family, Genus, and Species levels, would make the analysis more accurate and specific. In addition, it may indicate new relations between particular genera or species and the microbiome changes due to SARS-CoV-2 infection. This may be easier using deep learning and neural networks techniques, which would help indicate which taxa are relevant than others.

All codes and machine learning models are available on GitHub repository.

References

Aboulghate, M., Elaghoury, A., Elebrashy, I., Elkafrawy, N., Elshishiney, G., Abul-Magd, E., Bassiouny, E., Toaima, D., Elezbawy, B., Fasseeh, A., Abaza, S., & Vokó, Z. (2021). The Burden of Obesity in Egypt. *Frontiers in Public Health*, *9*. https://doi.org/10.3389/fpubh.2021.718978

Aguiar-Pulido, V., Huang, W., Suarez-Ulloa, V., Cickovski, T., Mathee, K., & Narasimhan, G. (2016). Metagenomics, metatranscriptomics, and metabolomics approaches for microbiome analysis: supplementary issue: bioinformatics methods and applications for big metagenomics data. *Evolutionary Bioinformatics*, *12*, EBO-S36436.

Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Xu, Z. Z., Kightley, E. P., Thompson, L. R., Hyde, E. R., & Gonzalez, A. (2017). Deblur rapidly resolves single-nucleotide community sequence patterns. *MSystems*, *2*(2).

Armougom, F., Henry, M., Vialettes, B., Raccah, D., & Raoult, D. (2009). Monitoring bacterial community of human gut microbiota reveals an increase in Lactobacillus in obese patients and Methanogens in anorexic patients. *PloS One*, *4*(9), e7125.

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., Fernandes, G. R., Tap, J., Bruls, T., & Batto, J. M. (2011). Erratum: Enterotypes of the human gut microbiome (Nature (2011) 473 (174-180)). *Nature*, *474*(7353).

Bäckhed, F., Ding, H., Wang, T., Hooper, L. V, Koh, G. Y., Nagy, A., Semenkovich, C. F., & Gordon, J. I. (2004). The gut microbiota as an environmental factor that regulates fat storage. *Proceedings of the National Academy of Sciences*, *101*(44), 15718–15723.

Bäckhed, F., Roswall, J., Peng, Y., Feng, Q., Jia, H., Kovatcheva-Datchary, P., Li, Y., Xia, Y., Xie, H., & Zhong, H. (2015). Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell Host & Microbe*, *17*(5), 690–703.

Barr, T., Sureshchandra, S., Ruegger, P., Zhang, J., Ma, W., Borneman, J., Grant, K., & Messaoudi, I. (2018). Concurrent gut transcriptome and microbiota profiling following chronic ethanol consumption in nonhuman primates. *Gut Microbes*, *9*(4), 338–356.

Baxter, N. T., Koumpouras, C. C., Rogers, M. A. M., Ruffin, M. T., & Schloss, P. D. (2016). DNA from fecal immunochemical test can replace stool for detection of colonic lesions using a microbiota-based model. *Microbiome*, *4*(1), 1–6.

Baxter, N. T., Ruffin, M. T., Rogers, M. A. M., & Schloss, P. D. (2016). Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome Medicine*, *8*(1), 1–10.

Beck, D., & Foster, J. A. (2015). Machine learning classifiers provide insight into the relationship between microbial communities and bacterial vaginosis. *BioData Mining*, *8*(1), 1–9.

Belzer, C., & De Vos, W. M. (2012). Microbes inside—from diversity to function: the case of Akkermansia. *The ISME Journal*, *6*(8), 1449–1458.

Berg Miller, M. E., Yeoman, C. J., Chia, N., Tringe, S. G., Angly, F. E., Edwards, R. A., Flint, H. J., Lamed, R., Bayer, E. A., & White, B. A. (2012). Phage–bacteria relationships and CRISPR elements revealed by a metagenomic survey of the rumen microbiome. *Environmental Microbiology*, *14*(1), 207–227.

Berger, A. K., Yi, H., Kearns, D. B., & Mainou, B. A. (2017). Bacteria and bacterial envelope components enhance mammalian reovirus thermostability. *PLoS Pathogens*, *13*(12), e1006768.

Bervoets, L., Van Hoorenbeeck, K., Kortleven, I., Van Noten, C., Hens, N., Vael, C., Goossens, H., Desager, K. N., & Vankerckhoven, V. (2013). Differences in gut microbiota composition between obese and lean children: a cross-sectional study. *Gut Pathogens*, *5*(1), 1–10.

Blaser, M. J., Chen, Y., & Reibman, J. (2008). Does Helicobacter pylori protect against asthma and allergy? *Gut*, *57*(5), 561–567.

Blaxter, M., Mann, J., Chapman, T., Thomas, F., Whitton, C., Floyd, R., & Abebe, E. (2005). Defining operational taxonomic units using DNA barcode data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1462), 1935–1943.

Blekhman, R., Goodrich, J. K., Huang, K., Sun, Q., Bukowski, R., Bell, J. T., Spector, T. D., Keinan, A., Ley, R. E., & Gevers, D. (2015). Host genetic variation impacts microbiome composition across human body sites. *Genome Biology*, *16*(1), 1–12.

Bokulich, N. A., Dillon, M. R., Bolyen, E., Kaehler, B. D., Huttley, G. A., & Caporaso, J. G. (2018). q2-sample-classifier: machine-learning tools for microbiome classification and regression. *Journal of Open Research Software*, *3*(30).

Bokulich, N. A., Kaehler, B. D., Rideout, J. R., Dillon, M., Bolyen, E., Knight, R., Huttley, G. A., & Caporaso, J. G. (2018). Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome*, *6*(1), 1–17.

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., & Asnicar, F. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, *37*(8), 852–857.

Boudewijns, M., Magerman, K., Verhaegen, J., Debrock, G., Peetermans, W. E., Donkersloot, P., Mewis, A., Peeters, V., Rummens, J. L., & Cartuyvels, R. (2003). Rothia dentocariosa, endocarditis and mycotic aneurysms: case report and review of the literature. *Clinical Microbiology and Infection*, *9*(3), 222–229.

Bradshaw, D. J., Marsh, P. D., Allison, C., & Schilling, K. M. (1996). Effect of oxygen, inoculum composition and flow rate on development of mixed-culture oral biofilms. *Microbiology*, *142*(3), 623–629.

Brestoff, J. R., & Artis, D. (2013). Commensal bacteria at the interface of host metabolism and the immune system. *Nature Immunology*, *14*(7), 676–684.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581–583.

Cani, P. D., Amar, J., Iglesias, M. A., Poggi, M., Knauf, C., Bastelica, D., Neyrinck, A. M., Fava, F., Tuohy, K. M., & Chabo, C. (2007). Metabolic endotoxemia initiates obesity and insulin resistance. *Diabetes*, *56*(7), 1761–1772.

Cani, P. D., Bibiloni, R., Knauf, C., Waget, A., Neyrinck, A. M., Delzenne, N. M., & Burcelin, R. (2008). Changes in gut microbiota control metabolic endotoxemia-induced inflammation in high-fat diet–induced obesity and diabetes in mice. *Diabetes*, *57*(6), 1470–1481.

Cani, P. D., Neyrinck, A. M., Fava, F., Knauf, C., Burcelin, R. G., Tuohy, K. M., Gibson, G. R., & Delzenne, N. M. (2007). Selective increases of bifidobacteria in gut microflora improve high-fat-diet-induced diabetes in mice through a mechanism associated with endotoxaemia. *Diabetologia*, *50*(11), 2374–2383.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Pena, A. G., Goodrich, J. K., & Gordon, J. I. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, *7*(5), 335–336.

Chambers, E. S., Viardot, A., Psichas, A., Morrison, D. J., Murphy, K. G., Zac-Varghese, S. E. K., MacDougall, K., Preston, T., Tedford, C., & Finlayson, G. S. (2015). Effects of targeted delivery of propionate to the human colon on appetite regulation, body weight maintenance and adiposity in overweight adults. *Gut*, *64*(11), 1744–1754.

Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., Qiu, Y., Wang, J., Liu, Y., & Wei, Y. (2020). Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *The Lancet*, *395*(10223), 507–513.

Cheung, K. S., Hung, I. F. N., Chan, P. P. Y., Lung, K. C., Tso, E., Liu, R., Ng, Y. Y., Chu, M. Y., Chung, T. W. H., & Tam, A. R. (2020). Gastrointestinal manifestations of SARS-CoV-2 infection and virus load in fecal samples from a Hong Kong cohort: systematic review and meta-analysis. *Gastroenterology*, *159*(1), 81–95.

Collado, M. C., Isolauri, E., Laitinen, K., & Salminen, S. (2008). Distinct composition of gut microbiota during pregnancy in overweight and normal-weight women. *The American Journal of Clinical Nutrition*, *88*(4), 894–899.

Dadkhah, E., Sikaroodi, M., Korman, L., Hardi, R., Baybick, J., Hanzel, D., Kuehn, G., Kuehn, T., & Gillevet, P. M. (2019). Gut microbiome identifies risk for colorectal polyps. *BMJ Open Gastroenterology*, *6*(1), e000297.

Dao, M. C., Everard, A., Aron-Wisnewsky, J., Sokolovska, N., Prifti, E., Verger, E. O., Kayser, B. D., Levenez, F., Chilloux, J., & Hoyles, L. (2016). Akkermansia muciniphila and improved metabolic health during a dietary intervention in obesity: relationship with gut microbiome richness and ecology. *Gut*, *65*(3), 426–436.

De Bandt, J.-P., Waligora-Dupriet, A.-J., & Butel, M.-J. (2011). Intestinal microbiota in inflammation and insulin resistance: relevance to humans. *Current Opinion in Clinical Nutrition & Metabolic Care*, *14*(4), 334–340.

de Macario, E. C., & Macario, A. J. L. (2009). Methanogenic archaea in health and disease: a novel paradigm of microbial pathogenesis. *International Journal of Medical Microbiology*, *299*(2), 99–108.

De Wit, N., Derrien, M., Bosch-Vermeulen, H., Oosterink, E., Keshtkar, S., Duval, C., de Vogel-van den Bosch, J., Kleerebezem, M., Müller, M., & van der Meer, R. (2012). Saturated fat stimulates obesity and hepatic steatosis and affects gut microbiota composition by an enhanced overflow of dietary fat to the distal intestine. *American Journal of Physiology-Gastrointestinal and Liver Physiology*, *303*(5), G589–G599.

Demigné, C., Morand, C., Levrat, M.-A., Besson, C., Moundras, C., & Rémésy, C. (1995). Effect of propionate on fatty acid and cholesterol synthesis and on acetate metabolism in isolated rat hepatocytes. *British Journal of Nutrition*, *74*(2), 209–219.

Den Besten, G., Van Eunen, K., Groen, A. K., Venema, K., Reijngoud, D.-J., & Bakker, B. M. (2013). The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism. *Journal of Lipid Research*, *54*(9), 2325–2340.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., & Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology*, *72*(7), 5069–5072.

Dollive, S., Peterfreund, G. L., Sherrill-Mix, S., Bittinger, K., Sinha, R., Hoffmann, C., Nabel, C. S., Hill, D. A., Artis, D., & Bachman, M. A. (2012). A tool kit for quantifying eukaryotic rRNA gene sequences from human microbiome samples. *Genome Biology*, *13*(7), 1–13.

Douglas, G. M., Hansen, R., Jones, C. M. A., Dunn, K. A., Comeau, A. M., Bielawski, J. P., Tayler, R., El-Omar, E. M., Russell, R. K., & Hold, G. L. (2018). Multi-omics differentially classify disease state and treatment outcome in pediatric Crohn's disease. *Microbiome*, *6*(1), 1–12.

Emoto, T., Yamashita, T., Sasaki, N., Hirota, Y., Hayashi, T., So, A., Kasahara, K., Yodoi, K., Matsumoto, T., & Mizoguchi, T. (2016). Analysis of gut microbiota in coronary artery disease patients: a possible link between gut microbiota and coronary artery disease. *Journal of Atherosclerosis and Thrombosis*, 32672.

Erickson, A. K., Jesudhasan, P. R., Mayer, M. J., Narbad, A., Winter, S. E., & Pfeiffer, J. K. (2018). Bacteria facilitate enteric virus co-infection of mammalian cells and promote genetic recombination. *Cell Host & Microbe*, *23*(1), 77–88.

Everard, A., Belzer, C., Geurts, L., Ouwerkerk, J. P., Druart, C., Bindels, L. B., Guiot, Y., Derrien, M., Muccioli, G. G., & Delzenne, N. M. (2013). Cross-talk between Akkermansia muciniphila and intestinal epithelium controls diet-induced obesity. *Proceedings of the National Academy of Sciences*, *110*(22), 9066–9071.

Everard, A., Lazarevic, V., Derrien, M., Girard, M., Muccioli, G. G., Neyrinck, A. M., Possemiers, S., Van Holle, A., François, P., & de Vos, W. M. (2011). Responses of gut microbiota and glucose and lipid metabolism to prebiotics in genetic obese and diet-induced leptin-resistant mice. *Diabetes*, *60*(11), 2775–2786.

Falkinham III, J. O. (2003). Mycobacterial aerosols and respiratory disease. *Emerging Infectious Diseases*, *9*(7), 763.

Fei, N., & Zhao, L. (2013). An opportunistic pathogen isolated from the gut of an obese human causes obesity in germfree mice. *The ISME Journal*, *7*(4), 880–884.

Flint, H. J., Scott, K. P., Louis, P., & Duncan, S. H. (2012). The role of the gut microbiota in nutrition and health. *Nature Reviews Gastroenterology & Hepatology*, *9*(10), 577.

Gallardo-Escárate, C., Valenzuela-Muñoz, V., Núñez-Acuña, G., Valenzuela-Miranda, D., Benaventel, B. P., Sáez-Vera, C., Urrutia, H., Novoa, B., Figueras, A., & Roberts, S. (2021). The wastewater microbiome: A novel insight for COVID-19 surveillance. *Science of The Total Environment*, *764*, 142867.

Gallopoulos, E., Houstis, E., & Rice, J. R. (1994). Computer as thinker/doer: Problem-solving environments for computational science. *IEEE Computational Science and Engineering*, *1*(2), 11–23.

Gao, X., Lin, S.-H., Ren, F., Li, J.-T., Chen, J.-J., Yao, C.-B., Yang, H.-B., Jiang, S.-X., Yan, G.-Q., & Wang, D. (2016). Acetate functions as an epigenetic metabolite to promote lipid synthesis under hypoxia. *Nature Communications*, *7*(1), 1–14.

Gao, Zhan, Tseng, C., Strober, B. E., Pei, Z., & Blaser, M. J. (2008). Substantial alterations of the cutaneous bacterial biota in psoriatic lesions. *PloS One*, *3*(7), e2719.

Gao, Zhanguo, Yin, J., Zhang, J., Ward, R. E., Martin, R. J., Lefevre, M., Cefalu, W. T., & Ye, J. (2009). Butyrate improves insulin sensitivity and increases energy expenditure in mice. *Diabetes*, *58*(7), 1509–1517.

Geurts, L., Neyrinck, A. M., Delzenne, N. M., Knauf, C., & Cani, P. D. (2014). Gut microbiota controls adipose tissue expansion, gut barrier and glucose metabolism: novel insights into molecular targets and interventions using prebiotics. *Beneficial Microbes*, *5*(1), 3–17.

Geva-Zatorsky, N., Sefik, E., Kua, L., Pasman, L., Tan, T. G., Ortiz-Lopez, A., Yanortsang, T. B., Yang, L., Jupp, R., & Mathis, D. (2017). Mining the human gut microbiota for immunomodulatory organisms. *Cell*, *168*(5), 928–943.

Gill, S. R., Pop, M., DeBoy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., Gordon, J. I., Relman, D. A., Fraser-Liggett, C. M., & Nelson, K. E. (2006). Metagenomic analysis of the human distal gut microbiome. *Science*, *312*(5778), 1355–1359.

Goecks, J., Nekrutenko, A., & Taylor, J. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, *11*(8), 1–13.

Greenhill, A. R., Tsuji, H., Ogata, K., Natsuhara, K., Morita, A., Soli, K., Larkins, J.-A., Tadokoro, K., Odani, S., & Baba, J. (2015). Characterization of the gut microbiota of Papua New Guineans using reverse transcription quantitative PCR. *PLoS One*, *10*(2), e0117427.

Gupta, V. K., Paul, S., & Dutta, C. (2017). Geography, ethnicity or subsistence-specific variations in human microbiome composition and diversity. *Frontiers in Microbiology*, *8*, 1162.

Hanada, S., Pirzadeh, M., Carver, K. Y., & Deng, J. C. (2018). Respiratory viral infection-induced microbiome alterations and secondary bacterial pneumonia. *Frontiers in Immunology*, *9*, 2640.

Hannigan, G. D., Duhaime, M. B., Ruffin, M. T., Koumpouras, C. C., & Schloss, P. D. (2017). Viral and bacterial communities of colorectal cancer. *BioRxiv*, 152868.

Hansen, L. B. S., Roager, H. M., Søndertoft, N. B., Gøbel, R. J., Kristensen, M., Vallès-Colomer, M., Vieira-Silva, S., Ibrügger, S., Lind, M. V, & Mærkedahl, R. B. (2018). A low-gluten diet induces changes in the intestinal microbiome of healthy Danish adults. *Nature Communications*, *9*(1), 1–13.

Hashimoto, T., Perlot, T., Rehman, A., Trichereau, J., Ishiguro, H., Paolino, M., Sigl, V., Hanada, T., Hanada, R., & Lipinski, S. (2012). ACE2 links amino acid malnutrition to microbial ecology and intestinal inflammation. *Nature*, *487*(7408), 477–481.

Hildebrandt, M. A., Hoffmann, C., Sherrill–Mix, S. A., Keilbaugh, S. A., Hamady, M., Chen, Y., Knight, R., Ahima, R. S., Bushman, F., & Wu, G. D. (2009). High-fat diet determines the composition of the murine gut microbiome independently of obesity. *Gastroenterology*, *137*(5), 1716–1724.

Hillmann, B., Al-Ghalith, G. A., Shields-Cutler, R. R., Zhu, Q., Gohl, D. M., Beckman, K. B., Knight, R., & Knights, D. (2018). Evaluating the information content of shallow shotgun metagenomics. *MSystems*, *3*(6).

Hoffman, J. I. E. (2019). Logistic regression. *Basic Biostatistics for Medical and Biomedical Practitioners*, 581–589.

Hong, C., Manimaran, S., Shen, Y., Perez-Rogers, J. F., Byrd, A. L., Castro-Nallar, E., Crandall, K. A., & Johnson, W. E. (2014). PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. *Microbiome*, *2*(1), 1–15.

Horz, H.-P., & Conrads, G. (2010). The discussion goes on: what is the role of Euryarchaeota in humans? *Archaea*, *2010*.

Hotamisligil, G. S. (2006). Inflammation and metabolic disorders. *Nature*, *444*(7121), 860–867.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., & Gu, X. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*, *395*(10223), 497–506.

Huse, S. M., Ye, Y., Zhou, Y., & Fodor, A. A. (2012). A core human microbiome as viewed through 16S rRNA sequence clusters. *PloS One*, *7*(6), e34242.

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., Creasy, H. H., Earl, A. M., FitzGerald, M. G., & Fulton, R. S. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, *486*(7402), 207.

Huttunen Maija-Riitta Hirvonen, Eila Iivanainen, Marja-Leena Katila, Kati, J. J. (2001). Comparison of mycobacteria-induced cytotoxicity and inflammatory responses in human and mouse cell lines. *Inhalation Toxicology*, *13*(11), 977–991.

Huttunen, K., Hyvärinen, A., Nevalainen, A., Komulainen, H., & Hirvonen, M.-R. (2003). Production of proinflammatory mediators by indoor air bacteria and fungal spores in mouse and human cell lines. *Environmental Health Perspectives*, *111*(1), 85–92.

Inc., A. (2020). *Conda — Conda documentation*. Anaconda Software Distribution. https://docs.conda.io/en/latest/

Janda, J. M., & Abbott, S. L. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *Journal of Clinical Microbiology*, *45*(9), 2761–2764.

Janssen, S., McDonald, D., Gonzalez, A., Navas-Molina, J. A., Jiang, L., Xu, Z. Z., Winker, K., Kado, D. M., Orwoll, E., & Manary, M. (2018). Phylogenetic placement of exact amplicon sequences improves associations with clinical information. *Msystems*, *3*(3).

Jie, Z., Xia, H., Zhong, S.-L., Feng, Q., Li, S., Liang, S., Zhong, H., Liu, Z., Gao, Y., & Zhao, H. (2017). The gut microbiome in atherosclerotic cardiovascular disease. *Nature Communications*, *8*(1), 1–12.

Johansson, M. E. V, Larsson, J. M. H., & Hansson, G. C. (2011). The two mucus layers of colon are organized by the MUC2 mucin, whereas the outer layer is a legislator of host–microbial interactions. *Proceedings of the National Academy of Sciences*, *108*(Supplement 1), 4659–4665.

Jumpertz, R., Le, D. S., Turnbaugh, P. J., Trinidad, C., Bogardus, C., Gordon, J. I., & Krakoff, J. (2011). Energy-balance studies reveal associations between gut microbes, caloric load, and nutrient absorption in humans. *The American Journal of Clinical Nutrition*, *94*(1), 58–65.

Jussila, J, Komulainen, H., Huttunen, K., Roponen, M., Hälinen, A., Hyvärinen, A., Kosma, V.-M., Pelkonen, J., & Hirvonen, M.-R. (2001). Inflammatory responses in mice after intratracheal instillation of spores of Streptomyces californicus isolated from indoor air of a moldy building. *Toxicology and Applied Pharmacology*, *171*(1), 61–69.

Jussila, J, Pelkonen, J., Kosma, V.-M., Mäki-Paakkanen, J., Komulainen, H., & Hirvonen, M.-R. (2003). Systemic immunoresponses in mice after repeated exposure of lungs to spores of Streptomyces californicus. *Clinical and Diagnostic Laboratory Immunology*, *10*(1), 30–37.

Jussila, Juha, Komulainen, H., Huttunen, K., Roponen, M., Iivanainen, E., Torkko, P., Kosma, V.-M., Pelkonen, J., & Hirvonen, M.-R. (2002). Mycobacterium terrae isolated from indoor air of a moisture-damaged building induces sustained biphasic inflammatory response in mouse lungs. *Environmental Health Perspectives*, *110*(11), 1119–1125.

Kapono, C. A., Morton, J. T., Bouslimani, A., Melnik, A. V, Orlinsky, K., Knaan, T. L., Garg, N., Vázquez-Baeza, Y., Protsyuk, I., & Janssen, S. (2018). Creating a 3D microbial and chemical snapshot of a human habitat. *Scientific Reports*, *8*(1), 1–12.

Karlsson, C. L. J., Önnerfält, J., Xu, J., Molin, G., Ahrné, S., & Thorngren-Jerneck, K. (2012). The microbiota of the gut in preschool children with normal and excessive body weight. *Obesity*, *20*(11), 2257–2261.

Khanna, S., & Tosh, P. K. (2014). A clinician's primer on the role of the microbiome in human health and disease. *Mayo Clinic Proceedings*, *89*(1), 107–114.

Khosravi, A., & Mazmanian, S. K. (2013). Disruption of the gut microbiome as a risk factor for microbial infections. *Current Opinion in Microbiology*, *16*(2), 221–227.

Kim, M., Lee, K.-H., Yoon, S.-W., Kim, B.-S., Chun, J., & Yi, H. (2013). Analytical tools and databases for metagenomics in the next-generation sequencing era. *Genomics & Informatics*, *11*(3), 102.

Koohi-Moghadam, M., Borad, M. J., Tran, N. L., Swanson, K. R., Boardman, L. A., Sun, H., & Wang, J. (2019). MetaMarker: a pipeline for de novo discovery of novel metagenomic biomarkers. *Bioinformatics*, *35*(19), 3812–3814.

Krajmalnik-Brown, R., Ilhan, Z., Kang, D., & DiBaise, J. K. (2012). Effects of gut microbes on nutrient absorption and energy regulation. *Nutrition in Clinical Practice*, *27*(2), 201–214.

Kunin, V., Engelbrektson, A., Ochman, H., & Hugenholtz, P. (2010). Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental Microbiology*, *12*(1), 118–123.

Lacey, J., & Crook, B. (1988). Fungal and actinomycete spores as pollutants of the workplace and occupational allergens. *The Annals of Occupational Hygiene*, *32*(4), 515–533.

Langille, M. G. I., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., Clemente, J. C., Burkepile, D. E., Thurber, R. L. V., & Knight, R. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology*, *31*(9), 814–821.

Laterza, L., Rizzatti, G., Gaetani, E., Chiusolo, P., & Gasbarrini, A. (2016). The gut microbiota and immune system relationship in human graft-versus-host disease. *Mediterranean Journal of Hematology and Infectious Diseases*, *8*(1).

LaValle, S. M., Branicky, M. S., & Lindemann, S. R. (2004). On the relationship between classical grid search and probabilistic roadmaps. *The International Journal of Robotics Research*, *23*(7–8), 673–692.

Leinonen, R., Sugawara, H., Shumway, M., & Collaboration, I. N. S. D. (2010). The sequence read archive. *Nucleic Acids Research*, *39*(suppl_1), D19–D21.

Ley, R. E., Bäckhed, F., Turnbaugh, P., Lozupone, C. A., Knight, R. D., & Gordon, J. I. (2005). Obesity alters gut microbial ecology. *Proceedings of the National Academy of Sciences*, *102*(31), 11070–11075.

Ley, R. E., Turnbaugh, P. J., Klein, S., & Gordon, J. I. (2006). Human gut microbes associated with obesity. *Nature*, *444*(7122), 1022–1023.

Li, Jing, Zhao, F., Wang, Y., Chen, J., Tao, J., Tian, G., Wu, S., Liu, W., Cui, Q., & Geng, B. (2017). Gut microbiota dysbiosis contributes to the development of hypertension. *Microbiome*, *5*(1), 1–19.

Li, Junhua, Jia, H., Cai, X., Zhong, H., Feng, Q., Sunagawa, S., Arumugam, M., Kultima, J. R., Prifti, E., & Nielsen, T. (2014). An integrated catalog of reference genes in the human gut microbiome. *Nature Biotechnology*, *32*(8), 834–841.

Li, N., Ma, W.-T., Pang, M., Fan, Q.-L., & Hua, J.-L. (2019). The commensal microbiota and viral infection: a comprehensive review. *Frontiers in Immunology*, *10*, 1551.

Li, S. S., Zhu, A., Benes, V., Costea, P. I., Hercog, R., Hildebrand, F., Huerta-Cepas, J., Nieuwdorp, M., Salojärvi, J., & Voigt, A. Y. (2016). Durable coexistence of donor and recipient strains after fecal microbiota transplantation. *Science*, *352*(6285), 586–589.

Liang, W., Feng, Z., Rao, S., Xiao, C., Xue, X., Lin, Z., Zhang, Q., & Qi, W. (2020). Diarrhoea may be underestimated: a missing link in 2019 novel coronavirus. *Gut*, *69*(6), 1141–1143.

Lopetuso, L. R., Ianiro, G., Scaldaferri, F., Cammarota, G., & Gasbarrini, A. (2016). Gut virome and inflammatory bowel disease. *Inflammatory Bowel Diseases*, *22*(7), 1708–1712.

Luckey, T. D. (1972). *Introduction to intestinal microecology*. Oxford University Press.

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *ArXiv Preprint ArXiv:1705.07874*.

Magne, F., Gotteland, M., Gauthier, L., Zazueta, A., Pesoa, S., Navarrete, P., & Balamurugan, R. (2020). The firmicutes/bacteroidetes ratio: a relevant marker of gut dysbiosis in obese patients? *Nutrients*, *12*(5), 1474.

Marcos-Zambrano, L. J., Karaduzovic-Hadziabdic, K., Loncar Turukalo, T., Przymus, P., Trajkovik, V., Aasmets, O., Berland, M., Gruca, A., Hasic, J., & Hron, K. (2021). Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment. *Frontiers in Microbiology*, *12*, 313.

Mardanov, A. V, Babykin, M. M., Beletsky, A. V, Grigoriev, A. I., Zinchenko, V. V, Kadnikov, V. V, Kirpichnikov, M. P., Mazur, A. M., Nedoluzhko, A. V, & Novikova, N. D. (2013). Metagenomic analysis of the dynamic changes in the gut microbiome of the participants of the MARS-500 experiment, simulating long term space flight. *Acta Naturae (Англоязычная Версия)*, *5*(3 (18)).

Martino, C., Kellman, B. P., Sandoval, D. R., Clausen, T. M., Marotz, C. A., Song, S. J., Wandro, S., Zaramela, L. S., Benítez, R. A. S., & Zhu, Q. (2020). Bacterial modification of the host glycosaminoglycan heparan sulfate modulates SARS-CoV-2 infectivity. *BioRxiv*.

McNeil, M. M., & Brown, J. M. (1994). The medically important aerobic actinomycetes: epidemiology and microbiology. *Clinical Microbiology Reviews*, *7*(3), 357–417.

Mossotto, E., Ashton, J. J., Coelho, T., Beattie, R. M., MacArthur, B. D., & Ennis, S. (2017). Classification of paediatric inflammatory bowel disease using machine learning. *Scientific Reports*, *7*(1), 1–10.

Neish, A. S. (2009). Microbes in gastrointestinal health and disease. *Gastroenterology*, *136*(1), 65–80.

O'Hara, A. M., & Shanahan, F. (2006). The gut flora as a forgotten organ. *EMBO Reports*, *7*(7), 688–693.

Obregon-Tito, A. J., Tito, R. Y., Metcalf, J., Sankaranarayanan, K., Clemente, J. C., Ursell, L.

K., Xu, Z. Z., Van Treuren, W., Knight, R., & Gaffney, P. M. (2015). Subsistence strategies in traditional societies distinguish gut microbiomes. *Nature Communications*, *6*(1), 1–9.

Onder, G., Rezza, G., & Brusaferro, S. (2020). Case-fatality rate and characteristics of patients dying in relation to COVID-19 in Italy. *Jama*, *323*(18), 1775–1776.

Pasolli, E., Schiffer, L., Manghi, P., Renson, A., Obenchain, V., Truong, D. T., Beghini, F., Malik, F., Ramos, M., & Dowd, J. B. (2017). Accessible, curated metagenomic data through ExperimentHub. *Nature Methods*, *14*(11), 1023.

Patel, J. B. (2001). 16S rRNA gene sequencing for bacterial pathogen identification in the clinical laboratory. *Molecular Diagnosis*, *6*(4), 313–321.

Pavlova, S. I., Wilkening, R. V, Federle, M. J., Lu, Y., Schwartz, J., & Tao, L. (2019). Streptococcus endopeptidases promote HPV infection in vitro. *Microbiologyopen*, *8*(1), e00628.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Perry, R. J., Peng, L., Barry, N. A., Cline, G. W., Zhang, D., Cardone, R. L., Petersen, K. F., Kibbey, R. G., Goodman, A. L., & Shulman, G. I. (2016). Acetate mediates a microbiome–brain–β-cell axis to promote metabolic syndrome. *Nature*, *534*(7606), 213–217.

Qin, N., Yang, F., Li, A., Prifti, E., Chen, Y., Shao, L., Guo, J., Le Chatelier, E., Yao, J., & Wu, L. (2014). Alterations of the human gut microbiome in liver cirrhosis. *Nature*, *513*(7516), 59–64.

Radwan, S., Gilfillan, D., Eklund, B., Radwan, H. M., El Menofy, N. G., Lee, J., Kapuscinski, M., & Abdo, Z. (2020). A comparative study of the gut microbiome in Egyptian patients with Type I and Type II diabetes. *PloS One*, *15*(9), e0238764.

Raymond, F., Ouameur, A. A., Déraspe, M., Iqbal, N., Gingras, H., Dridi, B., Leprohon, P., Plante, P.-L., Giroux, R., & Bérubé, È. (2016). The initial state of the human gut microbiome determines its reshaping by antibiotics. *The ISME Journal*, *10*(3), 707–720.

Rifkin, R., & Klautau, A. (2004). In defense of one-vs-all classification. *The Journal of Machine Learning Research*, *5*, 101–141.

Rinninella, E., Raoul, P., Cintoni, M., Franceschi, F., Miggiano, G. A. D., Gasbarrini, A., & Mele, M. C. (2019). What is the healthy gut microbiota composition? A changing ecosystem across age, environment, diet, and diseases. *Microorganisms*, *7*(1), 14.

Rizzatti, G., Lopetuso, L. R., Gibiino, G., Binda, C., & Gasbarrini, A. (2017). Proteobacteria: a common factor in human diseases. *BioMed Research International*, *2017*.

Robinson, C. M., Jesudhasan, P. R., & Pfeiffer, J. K. (2014). Bacterial lipopolysaccharide binding enhances virion stability and promotes environmental fitness of an enteric virus. *Cell Host & Microbe*, *15*(1), 36–46.

Roguet, A., Eren, A. M., Newton, R. J., & McLellan, S. L. (2018). Fecal source identification

using random forest. *Microbiome*, *6*(1), 1–15.

Säemann, M. D., Böhmig, G. A., Österreicher, C. H., Burtscher, H., Parolini, O., Diakos, C., Stöckl, J., Hörl, W. H., & Zlabinger, G. J. (2000). Anti-inflammatory effects of sodium butyrate on human monocytes: potent inhibition of IL-12 and up-regulation of IL-10 production. *The FASEB Journal*, *14*(15), 2380–2382.

Sahned, J., Saeed, D. M., & Misra, S. (2019). Sugar-free Workplace: A Step for Fighting Obesity. *Cureus*, *11*(12).

Salah, M., Azab, M., Ramadan, A., & Hanora, A. (2019). New insights on obesity and diabetes from gut microbiome alterations in Egyptian adults. *Omics: A Journal of Integrative Biology*, *23*(10), 477–485.

Sankaranarayanan, K., Ozga, A. T., Warinner, C., Tito, R. Y., Obregon-Tito, A. J., Xu, J., Gaffney, P. M., Jervis, L. L., Cox, D., & Stephens, L. (2015). Gut microbiome diversity among Cheyenne and Arapaho individuals from western Oklahoma. *Current Biology*, *25*(24), 3161–3169.

Santacruz, A., Collado, M. C., Garcia-Valdes, L., Segura, M. T., Martin-Lagos, J. A., Anjos, T., Marti-Romero, M., Lopez, R. M., Florido, J., & Campoy, C. (2010). Gut microbiota composition is associated with body weight, weight gain and biochemical parameters in pregnant women. *British Journal of Nutrition*, *104*(1), 83–92.

Schirmer, M., Smeekens, S. P., Vlamakis, H., Jaeger, M., Oosting, M., Franzosa, E. A., Ter Horst, R., Jansen, T., Jacobs, L., & Bonder, M. J. (2016). Linking the human gut microbiome to inflammatory cytokine production capacity. *Cell*, *167*(4), 1125–1136.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., & Robinson, C. J. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, *75*(23), 7537–7541.

Sedio, B. E., Boya P, C. A., & Rojas Echeverri, J. C. (2018). A protocol for high-throughput, untargeted forest community metabolomics using mass spectrometry molecular networks. *Applications in Plant Sciences*, *6*(3), e1033–e1033. https://doi.org/10.1002/aps3.1033

Shang, J., Ye, G., Shi, K., Wan, Y., Luo, C., Aihara, H., Geng, Q., Auerbach, A., & Li, F. (2020). Structural basis of receptor recognition by SARS-CoV-2. *Nature*, *581*(7807), 221–224.

Sia, S. F., Yan, L.-M., Chin, A. W. H., Fung, K., Choy, K.-T., Wong, A. Y. L., Kaewpreedee, P., Perera, R. A. P. M., Poon, L. L. M., & Nicholls, J. M. (2020). Pathogenesis and transmission of SARS-CoV-2 in golden hamsters. *Nature*, *583*(7818), 834–838.

Soenksen, L. R., Kassis, T., Conover, S. T., Marti-Fuster, B., Birkenfeld, J. S., Tucker-Schwartz, J., Naseem, A., Stavert, R. R., Kim, C. C., & Senna, M. M. (2021). Using deep learning for dermatologist-level detection of suspicious pigmented skin lesions from wide-field images. *Science Translational Medicine*, *13*(581).

Soliman, M. M., Ahmed, M. M., Salah-Eldin, A.-E., & Abdel-Aal, A. A.-A. (2011). Butyrate

regulates leptin expression through different signaling pathways in adipocytes. *Journal of Veterinary Science*, *12*(4), 319–323.

Stackebrandt, E., Rainey, F. A., & Ward-Rainey, N. L. (1997). Proposal for a new hierarchic classification system, Actinobacteria classis nov. *International Journal of Systematic and Evolutionary Microbiology*, *47*(2), 479–491.

Starkweather, J., & Moske, A. K. (2011). *Multinomial logistic regression*.

Sze, M. A., & Schloss, P. D. (2016). Looking for a signal in the noise: revisiting obesity and the microbiome. *MBio*, *7*(4).

Tashiro, M., Ciborowski, P., Klenk, H.-D., Pulverer, G., & Rott, R. (1987). Role of Staphylococcus protease in the development of influenza pneumonia. *Nature*, *325*(6104), 536–537.

Team, Rs. (2020). *RStudio: Integrated Development Environment for R* (1.3.1093). RStudio, PBC. http://www.rstudio.com/

Teixeira, T. F. S., Grześkowiak, Ł. M., Salminen, S., Laitinen, K., Bressan, J., & Peluzio, M. do C. G. (2013). Faecal levels of Bifidobacterium and Clostridium coccoides but not plasma lipopolysaccharide are inversely related to insulin and HOMA index in women. *Clinical Nutrition*, *32*(6), 1017–1022.

the SRA Toolkit Development Team. (n.d.). *SRA-Tools - NCBI*. Retrieved February 27, 2021, from http://ncbi.github.io/sra-tools/

Thomas, A. M., Manghi, P., Asnicar, F., Pasolli, E., Armanini, F., Zolfo, M., Beghini, F., Manara, S., Karcher, N., & Pozzi, C. (2019). Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nature Medicine*, *25*(4), 667–678.

Thursby, E., & Juge, N. (2017). Introduction to the human gut microbiota. *Biochemical Journal*, *474*(11), 1823–1836.

Topçuoğlu, B. D., Lesniak, N. A., Ruffin, M. T., Wiens, J., & Schloss, P. D. (2020). A framework for effective application of machine learning to microbiome-based classification problems. *Mbio*, *11*(3).

Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., & Segata, N. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods*, *12*(10), 902–903.

Turnbaugh, P. J., Bäckhed, F., Fulton, L., & Gordon, J. I. (2008). Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome. *Cell Host & Microbe*, *3*(4), 213–223.

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., Sogin, M. L., Jones, W. J., Roe, B. A., & Affourtit, J. P. (2009). A core gut microbiome in obese and lean twins. *Nature*, *457*(7228), 480–484.

Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., & Gordon, J. I.

(2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, *444*(7122), 1027–1031.

Tyakht, A. V, Kostryukova, E. S., Popenko, A. S., Belenikin, M. S., Pavlenko, A. V, Larin, A. K., Karpova, I. Y., Selezneva, O. V, Semashko, T. A., & Ospanova, E. A. (2013). Human gut microbiota community structures in urban and rural populations in Russia. *Nature Communications*, *4*(1), 1–9.

Valdes, A. M., Walter, J., Segal, E., & Spector, T. D. (2018). Role of the gut microbiota in nutrition and health. *Bmj*, *361*.

Vatanen, T., Kostic, A. D., d'Hennezel, E., Siljander, H., Franzosa, E. A., Yassour, M., Kolde, R., Vlamakis, H., Arthur, T. D., & Hämäläinen, A.-M. (2016). Variation in microbiome LPS immunogenicity contributes to autoimmunity in humans. *Cell*, *165*(4), 842–853.

Verberkmoes, N. C., Russell, A. L., Shah, M., Godzik, A., Rosenquist, M., Halfvarson, J., Lefsrud, M. G., Apajalahti, J., Tysk, C., & Hettich, R. L. (2009). Shotgun metaproteomics of the human distal gut microbiota. *The ISME Journal*, *3*(2), 179–189.

Walters, W. A., Xu, Z., & Knight, R. (2014). Meta-analyses of human gut microbes associated with obesity and IBD. *FEBS Letters*, *588*(22), 4223–4233.

Wang, J., Zhao, S., Liu, M., Zhao, Z., Xu, Y., Wang, P., Lin, M., Xu, Y., Huang, B., & Zuo, X. (2020). ACE2 expression by colonic epithelial cells is associated with viral infection, immunity and energy metabolism. *MedRxiv*.

Wang, M., Carver, J. J., Phelan, V. V, Sanchez, L. M., Garg, N., Peng, Y., Nguyen, D. D., Watrous, J., Kapono, C. A., & Luzzatto-Knaan, T. (2016). Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology*, *34*(8), 828–837.

Wirbel, J., Pyl, P. T., Kartal, E., Zych, K., Kashani, A., Milanese, A., Fleck, J. S., Voigt, A. Y., Palleja, A., & Ponnudurai, R. (2019). Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nature Medicine*, *25*(4), 679–689.

Wölfel, R., Corman, V. M., Guggemos, W., Seilmaier, M., Zange, S., Müller, M. A., Niemeyer, D., Jones, T. C., Vollmar, P., & Rothe, C. (2020). Virological assessment of hospitalized patients with COVID-2019. *Nature*, *581*(7809), 465–469.

Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S. A., Bewtra, M., Knights, D., Walters, W. A., & Knight, R. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science*, *334*(6052), 105–108.

Wu, H., Cai, L., Li, D., Wang, X., Zhao, S., Zou, F., & Zhou, K. (2018). Metagenomics biomarkers selected for prediction of three different diseases in Chinese population. *BioMed Research International*, *2018*.

Xiao, F., Tang, M., Zheng, X., Liu, Y., Li, X., & Shan, H. (2020). Evidence for Gastrointestinal Infection of SARS-CoV-2. *Gastroenterology*, *158*(6), 1831-1833.e3. https://doi.org/10.1053/j.gastro.2020.02.055

Xu, P., Li, M., Zhang, J., & Zhang, T. (2012). Correlation of intestinal microbiota with overweight and obesity in Kazakh school children. *BMC Microbiology*, *12*(1), 1–6.

Xu, R., Lu, R., Zhang, T., Wu, Q., Cai, W., Han, X., Wan, Z., Jin, X., Zhang, Z., & Zhang, C. (2020). *Temporal dynamics of human respiratory and gut microbiomes during the course of COVID-19 in adults*.

Xu, Y., Li, X., Zhu, B., Liang, H., Fang, C., Gong, Y., Guo, Q., Sun, X., Zhao, D., & Shen, J. (2020). Characteristics of pediatric SARS-CoV-2 infection and potential evidence for persistent fecal viral shedding. *Nature Medicine*, *26*(4), 502–505.

Yang, T., Santisteban, M. M., Rodriguez, V., Li, E., Ahmari, N., Carvajal, J. M., Zadeh, M., Gong, M., Qi, Y., & Zubcevic, J. (2015). Gut dysbiosis is linked to hypertension. *Hypertension*, *65*(6), 1331–1340.

Ye, Z., Zhang, N., Wu, C., Zhang, X., Wang, Q., Huang, X., Du, L., Cao, Q., Tang, J., & Zhou, C. (2018). A metagenomic study of the gut microbiome in Behcet's disease. *Microbiome*, *6*(1), 1–13.

Yildiz, S., Mazel-Sanchez, B., Kandasamy, M., Manicassamy, B., & Schmolke, M. (2018). Influenza A virus infection impacts systemic microbiota dynamics and causes quantitative enteric dysbiosis. *Microbiome*, *6*(1), 1–17.

Zackular, J. P., Rogers, M. A. M., Ruffin, M. T., & Schloss, P. D. (2014). The human gut microbiome as a screening tool for colorectal cancer. *Cancer Prevention Research*, *7*(11), 1112–1121.

Zaura, E., Keijser, B. J. F., Huse, S. M., & Crielaard, W. (2009). Defining the healthy" core microbiome" of oral microbial communities. *BMC Microbiology*, *9*(1), 1–12.

Zeller, G., Tap, J., Voigt, A. Y., Sunagawa, S., Kultima, J. R., Costea, P. I., Amiot, A., Böhm, J., Brunetti, F., & Habermann, N. (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular Systems Biology*, *10*(11), 766.

Zhernakova, A, Kurilshikov, A., Bonder, M. J., Tigchelaar, E. F., Schirmer, M., Vatanen, T., Mujagic, Z., Vila, A. V, Falony, G., & Vieira-Silva, S. (2016). Lifelines cohort studyand Weersma RK. *Feskens EJM, Netea1 MG, Gevers D, Jonkers D, Franke L, Aulchenko YS, Huttenhower C, Raes J, Hofker MH, Xavier RJ, Wijmenga C, Fu J*, 565–569.

Zhernakova, Alexandra, Kurilshikov, A., Bonder, M. J., Tigchelaar, E. F., Schirmer, M., Vatanen, T., Mujagic, Z., Vila, A. V., Falony, G., & Vieira-Silva, S. (2016). Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science*, *352*(6285), 565–569.

Zimmer, J., Lange, B., Frick, J.-S., Sauer, H., Zimmermann, K., Schwiertz, A., Rusch, K., Klosterhalfen, S., & Enck, P. (2012). A vegan or vegetarian diet substantially alters the human colonic faecal microbiota. *European Journal of Clinical Nutrition*, *66*(1), 53–60.

Zou, Y., Ju, X., Chen, W., Yuan, J., Wang, Z., Aluko, R. E., & He, R. (2020). Rice bran attenuated obesity via alleviating dyslipidemia, browning of white adipocytes and modulating gut microbiota in high-fat diet-induced obese mice. *Food & Function*, *11*(3), 2406–2417.

Zuo, T., Zhang, F., Lui, G. C. Y., Yeoh, Y. K., Li, A. Y. L., Zhan, H., Wan, Y., Chung, A. C. K., Cheung, C. P., & Chen, N. (2020). Alterations in gut microbiota of patients with COVID-19 during time of hospitalization. *Gastroenterology*, *159*(3), 944–955.

Ahmed Adel Aboushanab

E-mail: ahmedadelaboushanab@aucegypt.edu, a.adel@proteinea.com
Phone: +201143450774
Address: 4, Zahraa Plaza, Zahraa Nasr City, Nasr City, Cairo, Egypt
LinkedIn: https://www.linkedin.com/in/ahmed-aboushanab-13b029a1/
ResearchGate: https://www.researchgate.net/profile/Ahmed_Aboushanab
GitHUB: https://github.com/AhmedAboushanab

Education

- Biotechnology Masters Student at the American University in Cairo (AUC). **Sep 2018 – Jan 2022**
- Scored 311/340 in **GRE.** **Apr 2018**
- Bachelor of Science (Biomedical Sciences major; Molecular and Cell Biology concentration) from The University of Science and Technology - Zewail City. **2013 – 2017**

Work experience

- Research and Development Lead at Proteinea. **Jun 2019 - present**
- Graduated from Indiebio Accelerator Program. **Feb 2021 – Jul 2021**
- Teaching assistant at the American University in Cairo (AUC)- Biology department. **Sep 2020 – Jun 2021**
- Graduated from Changelabs Accelerator Program. **Jul 2019 – Aug 2019**
- Awarded an Internship at Molecular Biology and Virology Lab (MBVL) at Helmy Institutes for Medical Sciences (HIMS) with responsibility for **2015 – 2018** **2015 – 2018**
  - Cloning (Unknown and EGFP) genes in Baculovirus vector.
  - Preparation of research plans and ideas.

- Attended the 3rd Middle East Molecular Biology Congress & Exhibition in Doha-Research Complex / Qatar University. **2016**
- Conference oral presentation about a biomimetic microfluidic device - 2nd IEEE EMBS International Student Conference (ISCEGYPT 2015) at Cairo University**.** **2015**
- Completed the NIH Web-based training course "Protecting Human Research Participants". **2015**

Publications

- 'Pancreas on a chip' abstract presentation at: IEEE EMBS International Students Conference (2015) at Cairo University.

Skills

- Excellent in Microsoft Office (Word, Excel, Power point).
- Fluent in English and Arabic.
- Python Programming.
- R programming.
- Gene Cloning.

Extracurricular activities

- Participated as student leader for the new graduate student orientation. **Jan 27th 2019**
- Coauthor of the book (Zewail..The City) (المدينة زويل).
- Best member in the organizing committee in Zewail City Conference and Exhibition on Biomedical Sciences (ZCEBS). **2017**

REFERENCES FURNISHED UPON REQUEST