Development of pipeline for gut microbiome data analysis using R programming

Binu Mathew
Master's Thesis
Master's Degree Programme in Digital Health and Life Sciences
Department of Computing
University of Turku
January 2022

ABSTRACT

UNIVERSITY OF TURKU
Department of Computing/Faculty of Technology

MATHEW, BINU: Development of pipeline for gut microbiome data analysis using R programming

Master's Thesis, 53 p
Bioinformatics
January 2022

Microorganisms are present everywhere in the environment. They are present on and inside of all higher organisms and they can be harmful as well as useful for the host. Microorganisms play an important role in the evolution and thus make it important to study them. Microbiota is the name given for the totality of micro-organisms present in the environment and the microbiota of a particular habitat is known as the microbiome. Due to its presence in the biosphere and the human body, a large number of experiments including the human microbiome project (HMP) are carried out and a huge amount of data is produced.

The gut microbiome is one of the most studied microbiota. Microorganisms living in the host gut play several important roles like metabolic activity, anti-cancer activity, and anti-infection activity. A healthy individual has a balanced microbiota and any imbalance in them causes various diseases and health issues. Obesity or even certain cancers are caused by the imbalance of the microbes in the intestine.

This project focuses on developing a bioinformatics pipeline to analyse the gut microbiome in a dietary intervention study. The data used for this project was collected from a study which was aimed to study gastric cancer where we were exploring the role and difference in composition of the gut microbiome. This thesis handles the entirety of the data processing and analysis through R involving: pre-processing of data; phyloseq object generation; statistical analysis and visualization; diversity indices calculation (alpha and beta) and composition analysis of microbiomes. After completion of the project, one would be able to analyse the gut microbiome data with minimal effort.

Keywords
16S rRNA sequencing, OTU, gut microbiome, *phyloseq*, alpha diversity, beta diversity, ALDEx2, Gastric adenocarcinoma, GIST

ACKNOWLEDGEMENT

I am thankful to my supervisor, Assoc. Prof. Leo Lahti, for allowing me to do a thesis under his supervision. I especially thank him for his time, advice, and patience. I would also like to extend my sincere thanks to Dr. Virinder Sarhadi and Prof. Sakari Knuutila for providing us with the data for benchmarking the pipeline. A special thanks to Wisam Saleem for providing ideas on troubleshooting coding errors.

I am grateful to my teachers, Dr. Juho Heimonen, Dr. Martti Tolvanen, Assoc. Prof. Filip Ginter and Adj. Prof. Csaba Ortutay, for their excellent lectures, guidance, and support during my studies. Special thanks to Juho and Martti for helping in other study program-related matters.

Last but not least, I thank my family and friends for their motivation to complete my studies. I am also thankful for my family's patience and support throughout my studies.

TABLE OF CONTENTS

1. INTRODUCTION

All multicellular organisms harbour diverse communities of microorganisms or microbes, collectively called microbiota. In this view, multicellular organism is considered as a holobiont, a combined entity of host organism and its associated microbial community (1). The microbial community serves different life-functions aiding in the survival of the host. Humans are no different and the number of microorganisms living on and inside humans are estimated to be 1. 3x times more than the somatic and germ cells of the host (2). In fact, the collection of genes of the microbial community known as the microbiome is considered as our second genome (3). An upsurge in human microbiome research in the past two decades has opened new vistas to many unanswered questions particularly pertaining to human health and diseases.

Bioinformatics is one area of study leading to a greater understanding of the microbiome. In the field of microbiology, bioinformatics has become an essential tool. The use of bioinformatics is inherent to virtually every modern research project in biology, whether it is analyzing DNA or protein sequences or parsing the information in massive gigabyte-sized data sets. A particularly well-known example is the next-generation sequencing (NGS), which transformed fields such as molecular systems, population genetics, quantitative genetics and microbial ecology (4). Researchers can understand the microbiome's composition and metabolic activities by conducting data meta-analyses based using bioinformatics. The gut microbiome has a high density of microbes but has a relatively low proportion of culturable bacteria (5). Microbiomes were studied using individual bacteria or microbial communities in co-cultures in the early phases of the microbiome study. NGS made it possible to accurately identify most members of a complex microbial community. High-throughput sequencing technologies and bioinformatics analysis have revolutionized gut microbiome composition and function research (6).

Bioinformatics tools help characterize the microbiome's impact on the evolution of a disease or the effect of an illness or intervention on the microbiome. When combined with statistical analysis, rapidly evolving bioinformatics tools provide insight into the association of microbiome with diseases (7). For these purposes, it is essential to develop a pipeline to analyze

the human gut microbiome which could shed light on the association between gut microbiome and diseases by comparing the gut microbiome of healthy people with the patients.

Therefore, this thesis explores the relationship between the gut microbiome and gastric cancer by developing a pipeline to analyze the human gut microbiome. In this study, the gut microbiome of healthy people is compared with the gut microbiome of gastric cancer patients.

2. LITERATURE REVIEW

2. 1 HUMAN MICROBIOME

The term 'human microbiome' was introduced by Joshua Lederberg in 2001 (8). Since then, research has been carried out to find the holistic role of microbes in humans, their impacts on human health and what specific functions they have in each habitat, or more precisely in each organ system. The main habitats are mouth, oesophagus, stomach, small intestine, large intestine, ear canal, nasopharynx, oropharynx, skin, penis and vagina (Figure 1). The composition of dominant bacterial phyla also varies according to body habitats (9). The main functions contributed by the human microbiome are: maintains a healthy digestive tract, helps digestion, immunity against pathogens, regulates cardiovascular system, anti-inflammatory and anti-oxidant activities, supplement metabolic potential and maintains mental health (10). Dysbiosis in human microbiome lead to various diseases and some of the leading factors involved in altering microbiome composition are listed Table 1.

Current knowledge on human microbiome is well advanced that not only the scientific world but also the common mass anticipates for the next step to understand how the human microbiome influences the digestive system or mental health. A human's microbiome is formed from its birth and is highly influenced by the mode of delivery and feeding (11). Microbiota acquired during the first few years of life play a crucial role in maintaining health as an adult. The microbiota composition begins to be similar to adult in the first year of life and the phylogenetic diversity steadily increases over time (12). Individual microbiomes vary in composition (13); human genomes have a limited role and many of the characteristics of the microbial community are influenced by the environment. (14). This again brings in the fact that although microbiome is shaped according to the tissues where they reside, each human follows their own specific microbial composition in each tissue. The skin, for instance, has a unique microbial composition that is unique to each individual (15). In the same manner, despite long-term physical oral interaction between humans influencing the composition of microbial communities over time, the oral microbiome has not unified in humans. (16). Hence, a healthy microbiome cannot be given a general definition, rather it is highly influenced by many factors and should be considered personalized to each individual.

A human body consists of 1.3x as many microbial cells as our own (2). Our understanding of a 'healthy microbiome' concept is evolving as well identifying the structure and composition of microbial community associated with each tissue or organ in the human body. Initial research to understand the ecology of healthy microbiomes have been to identify 'core' taxa that can be related to healthy individuals in contrast to diseased individuals that lack them. However, several studies on ecological diversity in healthy individuals showed that the taxonomy of microbial community composition varied widely and hence determining 'core' taxa remains irrelevant (17).

Regardless, from the observations from several studies, a recent article suggests five broad definitions on core microbiome – temporal, ecological, host-adapted, functional, and common core. Temporal, ecological, and common core relates to patterns in spatial, temporal and ecological microbiome dynamics, while the functional and host-adapted cores are related to the function of the host and its fitness (18). Therefore, with deeper investigation into human microbiome is widening our knowledge on how exactly it promotes human health.

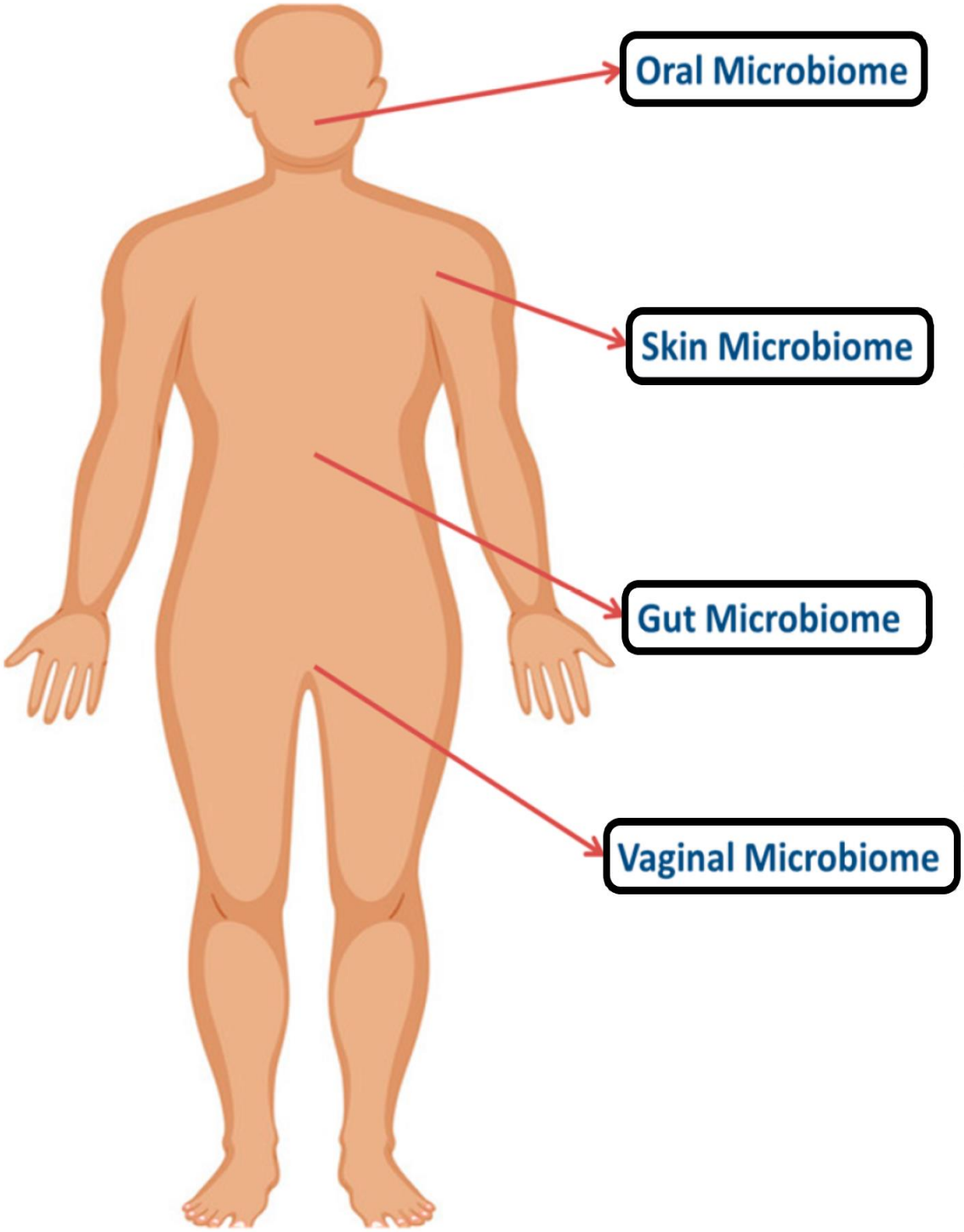


Figure 1: Different habitat of human microbiota. Figure modified from (19). The figure is licensed under CC BY 4. 0.

Table 1: Factors affecting the dysbiosis of human gut microbiome.

Causes	Factors	Effects	References
Environmental factors	Ethnicity and geographical location	Variations in gut microbiome among different populations, affects neurologic disorders, diabetes and obesity	(20–22)
Exposure to pathogens and allergens	Food intolerance	Dysbiosis in the population of microbially responsive RORγt-positive FOXp3-positive regulatory T cells causes food intolerance	(23)
	Low microbial exposure during pregnancy period	Allergic diseases such as atopic eczema, food allergy, asthma	(24)
Medicine and antibiotic use	Antimicrobial drug use	Decrease in microbial species diversity in gut environment.	(25)
Exposure to pollutants	Air pollution, ozone exposure	Alterations in composition and function of gut microbiome	(26,27)
Diet	Diet	Diet mainly affects the gut and salivary microbiome and results in oral diseases, obesity, Autoimmune diseases, host metabolism and many chronic diseases	(28–31)
Intoxication	Alcohol	Affects gut microbiome composition and metabolism contributing to alcohol-induced oxidative stress and subsequently developing alcoholic liver disease (ALD)	(32)
	Smoking	Reduced gut microbiome diversity, increasing oxidative stress and changes in acid- base balance	(33)

2. 2 GUT MICROBIOME

Microbes residing in the human gut and their genes are collectively known as the gut microbiome and has a major role in behaviour of host, metabolism, nutrient absorption, development and immunity (34,35). The disruption of microbial composition known as dysbiosis has been associated with diabetes, obesity, inflammatory bowel diseases and autoimmune diseases (Table 2). An individual acquires gut microbiome from birth and linearly increases during the first three years of life. Several factors such as host genotype, age, sex, diet and disease determine gut microbiome composition (36). The adult gut microbiome is relatively stable and the dominant bacterial phyla include Bacteroidetes, Firmicutes and Actinobacteria (37).

A huge understanding of human microbiome comes from research related to digestive system or the gastrointestinal tract. Most studies are contributed by investigating the structure and composition of microbiome in the intestinal tract or gut contributed by a vast diversity of microbes (38). It has been demonstrated that the human microbiome contains 3. 3 million unique genes, 150 times more than the sequence of human genome. A bacterial diversity analysis has revealed that about 1000 species of bacteria are found in the human gut, and most of these belong to Firmicutes and Bacteroidetes. Furthermore, most individuals share 50–100 bacterial species regardless of the frequency of occurrence at the phylotype level, and the majority of human gut samples have shown to harbour as many as 6000 functional gene groups. Animal studies indicate that gut bacteria are essential for controlling gut metabolism and in contributing to the health of the host immune system (39). The interest in research of gut microbiome arose and is continuing from the fact it could be linked to several survival functions such as digestion and providing immunity against pathogens as well as several dysbiosis conditions such as obesity, diabetes, several inflammatory bowel diseases (IBD) and autoimmune diseases (37). Although host genotype, age and sex are factors affecting the gut microbiome, diet is the most influential factor on shaping gut microbiota. Thus, the gut microbiota is considered as a vital partner of human cells interacting with virtually all human cells (40).

Most studies related to gut bacteria have been on Western populations in USA, Europe and Canada. These studies are now being expanded into non-western diets in order to investigate the impact of diet variation on gut microbial community. An interesting study found

that the gut bacteria *Bacteroides plebeius* carry a unique gene coding for the enzyme porphyranase that is unique to the Japanese population. It is speculated to have transferred from the marine *Bacteroides* spp. It has the capability to degrade seaweed and hence helps in its digestion in Japanese individuals who consume this seaweed in their diet (41). Research on the influence of the diet on the gut microbiota is fascinating since it has been identified they are both complementary. Modulating diet imposes change in gut microbiota and vice versa. New therapeutic approaches may come from this avenue of research (42). Another major influential factor on shaping human microbiome is lifestyle. Owning or close association with animals, physical interaction, exercise, sleep patterns, stress and occupation-related lifestyles are all some of the factors coming under lifestyle (14).

The diversity of the human intestinal microbiota is very vast as it contains 1. 3x more cells than the body cells (2). The human gut microbiota contains 80% uncultivated and 60% new species (39). A total of 70 Bacterial divisions have been described so far, and thirteen Archaeal divisions, but the majority are Firmicutes and Bacteroidetes, i.e., of the 395 bacteria, 301 are Firmicutes, 65 are Bacteroidetes, and the remaining 29 are scattered among eight other divisions (43–45). In one study, gut microbiome sequence data for 124 Europeans generated 576.7 Gb of data. More than 3.3 million unique bacteria genes were assembled and characterized from it, which is 150 times as many as the human genome (46). From the 'core gut microbiome'; a total of 6313 functional orthologous groups were found (39). Firmicutes, Actinobacteria, Proteobacteria, Bacteroidetes, and Fusobacteria are the most prevalent phyla in the oral cavity, accounting for more than 99% of all phyla and SR1, TM7, Tenericutes, Cyanobacteria, Synergistetes, and Spirochaetes are rare ones (47). The esophagus showed the presence Actinobacteria, Bacteroides, Firmicutes, Fusobacteria, Proteobacteria, and TM7. It was found that Prevotella, Veillonella, and Streptococcus were the most prevalent bacteria, and the distal esophagus community was similar to that of the oral microbiota (48). During study of the gastric tract, it was found that there are diverse groups of gastric microbes, mainly from Proteobacteria, Firmicutes, Actinobacteria, Bacteroidetes, and Fusobacteria (49). As compared to the oral and esophageal microbial communities, this community was quite different. This is just a brief account to visualize the vast amount of data these microbiomes can provide us and the much more enormous information they will provide by their linkage to bodily functions. The three major questions in gut microbiome research to be answered are to: (1) characterize the microbial communities in different sites of the human body (e.g., skin, mouth, gut etc.), (2)

to determine the core microbiome of individuals based on health conditions, location, or ethnicity and (3) to study the relationship between changes in microbiome and disease (50).

Knowledge about human microbiomes, especially gut microbiome has reached a critical inflection point. From descriptive studies on the composition of gut microbiome attempts are raving to associate them to functions and diseases, thus taking a clinical approach to find preventive or therapeutic measures (14). One of the major branches of research studied is the relation between gut microbiome and gastric cancer.

Table 2: Major gut microbiome related diseases summarized from other studies.

DISEASE	ASSOCIATION	REFERENCE
Crohn's disease, liver cirrhosis, liver cancer, irritable bowel syndrome, obesity	Imbalance in bile acid production by gut microbes	(51)
Infant asthma	Growth of Enterobacteriaceae	(52)
	Reduced relative abundance of bacterial genera *Rothia*, *Veillonella*, *Faecalibacterium* and *Lachnospira*	(53)
Obesity	Disruption of gastrointestinal microbiota impacts insulin resistance, inflammation and adiposity via interactions with epithelial and endocrine cells	(54)
	Decrease of relative proportion of Bacteroidetes	(55)
Gastritis, hypochlorhydria, duodenal ulcers, peptic ulcer and gastric cancer	*Helicobacter pylori* pathogenesis	(56)
Parkinson's disease	Dysbiosis of gut microbiome may be a contributing factor causing loss of dopaminergic neurons	(57)

DISEASE	ASSOCIATION	REFERENCE
Ulcerative colitis in children	Decreased diversity in gut microbiome	(58)
Type 2 diabetes	Altered overall composition of bacterial community, for instance, reduction in butyrate-producing bacteria and Akkermansia municiphila.	(59)
Autoimmune diseases	Caused by leaky gut as it causes deviation in gut microbiota leading to increased intestinal permeability.	(60)
Dementia	Dysfunction in the brain-gut microbiota axis due to intervention of factors such as diet, stress-hormones etc.	(61)
Celiac disease (CD)	Higher proportion of Bacterioides in CD patients compared to controls	(62,63)
	Higher amounts of Proteobacteria	(64,65)

2. 3 GASTRIC CANCER AND ROLE OF GUT MICROBIOME

Disruption in microbial composition shifts the health of the host to a diseased state causing serious ailments such as cancer. Gastric cancer is one of the common types of cancer and the second leading cause of cancer death in humans (66). Studies show that the main causes of gastric cancer are diet, tobacco, infection with *Helicobacter pylori (H. pylori)* bacteria and familial history (67). Gastric inflammation and carcinogenesis are strongly associated with *H. pylori* infection. It has been recognized as a carcinogen in 1994, although specific mechanism of its action in causing cancer is not known until date. However, countries with high incidence of gastric cancers have high prevalence of *H. pylori* infection and decline in infection decreased gastric cancer rates (67).

Few researches on the gut microbiome in gastric cancers focus on *H. pylori* and how its population affects other members of the gut microbiome. Individuals with *H. pylori* infection tended to decrease abundance of healthy gut microbiome community members of Bacteroidetes, Firmicutes and Actinobacteria (68). But further studies are required to find the exact mechanism to understand the connection of *H. pylori* and how it affects the composition of gut microbiome, how its impact on other members of the gut microbiome leads to gastric cancer. Hence, an account on the composition of gut microbial community could provide new insights into diagnosis and treatments of gastric cancer.

Gastric cancer is fifth most common cancers in world with an estimated more than one million new cases per year. An estimated 769 000 patients died from gastric cancer in 2020 (69). The main reasons for gastric cancer are aetiological factors, diet, use of tobacco, familial history of gastric cancer and infection with bacterium *H. pylori*. *H. pylori* infection is strongly linked to gastric inflammation and cancer (70,71). The most common type of gastric cancer is gastric adenocarcinoma. According to Lauren's classification gastric cancer which is histologically grouped, into diffuse and intestinal subtypes, but some cases can exhibit both diffuse and intestinal features, denoted as 'indeterminate' or 'mixed' phenotypes (72). *H. pylori* infection is associated with both subtypes of gastric cancer, but it has more prevalence in the intestinal subtype (73). Both subtypes, however, have various carcinogenic pathways and pathogenesis. The development of intestinal adenocarcinoma is generally associated with stomach inflammation is preceded by numerous premalignant stages, like intestinal metaplasia. On the other hand, diffuse adenocarcinomas have poor cell differentiation often results in poor

outcome and survival compared to intestinal adenocarcinomas (74). Adenocarcinomas of the intestinal tract have higher levels of genetic imbalance which includes microsatellite and chromosome instability (75). Mutations of the E-cadherin gene (CDH1) lead to diffuse adenocarcinomas, which are genetically more stable (76). GIST (gastrointestinal stromal tumors) are very rare and distinct from gastric adenocarcinomas. Usually, they are mainly found in the stomach, but they are also present elsewhere in the gastrointestinal tract. The tumors arise from stromal cells and carry mutations in PDGRA or KIT (77). Patients with gastric adenocarcinomas exhibit altered microbiota (78), although there is very little information available regarding alterations to intestinal microbiota in patients with gastric adenocarcinoma and GIST.

2. 4 ROLE OF BIOINFORMATICS IN ANALYSING MICROBIOME

2.4.1 EARLY BIOINFORMATICS APPROACHES

Human genome project helped us advance in understanding structure, function and genetics of the human race. However, it was later comprehended that critical functions related to the survival of the human host is possible only with the associated microbial community. After the completion of Human Genome Project, the study of human biology requires more than decoding the human genome, according to Julian Davies (79). More than 1,000 bacterial species dwells on and in the human body, with a serious impact on life. Despite the 30,000 genes in the human genome, he suggested that these bacteria might harbor 2 to 4 million unidentified genes, and these two sets of genes might determine the health of the individual. (79). A complex, body-habitat specific microbial ecosystem has evolved as human beings have co-evolved with trillions of microorganisms living on or inside their bodies (80). A second human genome project to conduct a complete metagenome study on the bacterial communities present in skin, reproductive tract, intestinal tract, and mouth was called for in 2001 (81). An extensive analysis of the human microbiome was not started until 2007 despite its importance being overlooked in the massive limelight of analysis of the human genome (39).

From the 1970s, isolation of human gut microbiota was by culturing samples in suitable growth media in laboratory aseptic conditions. However, it was possible to identify only around 400 – 500 isolates. But with the advent of sequencing techniques and later with high-throughput sequencing it is now feasible to identify even strains that cannot be cultured in lab conditions.

Although several techniques have been developed nowadays the two main approaches in human microbiome studies are 16S rRNA gene targeted sequencing and whole genome sequencing. Most studies still favor the 16S rRNA gene sequencing over whole genome sequencing due to technical and economic constraints although it may compromise the amount of sequence information that can be retrieved. Once the sequences are obtained the data is analyzed statistically to interpret a biologically meaningful information (50).

2.4.2 CURRENT MICROBIOME ANALYSIS TECHNIQUES

Designing and developing a comprehensive plan from raw data to final analysis and reporting depends on the several methods and concepts available in microbial ecology (82). Bioinformatics, nowadays more precisely defined for the tools used to gather information from sequencing data and provide statistical analysis on them are still novel to hardcore lab researchers. It is very perplexing for a biologist to be lost in the vast expansion of bioinformatics measures that can be used to manipulate sequence data. Therefore, designing a bioinformatics pipeline helps them through each and every step guiding on what to do next with the huge amounts of sequence data. However, it has been understood that a multi-disciplinary approach to answer a research question is most essential. Well-designed workflows involving best practices can supplement such an approach. It should involve accurate examples and guidance for the user to choose the methods not only maintaining flexibility but also providing possibility of customizing the workflow (82).

Microbiome research involves data sets such as counts of taxonomic units, genes, or metabolites that are typical to this field of research. This information is comprehensively appreciated when complemented with other information such as taxonomic classifications, phylogenies and nucleotide sequences. Data packages can complement the algorithmic R packages. It may be larger than the standard algorithm packages. One of the advantages of data packages is that they come with well-documented model data sets facilitating the method development, unit tests, and tutorials. Although R data packages have long been used in bioinformatics, lately these data packages are widely utilized in the microbiome field as well. They are used to provide data from recent microbiome studies at taxonomic and functional levels (82).

The 16S rRNA gene sequencing is the most common method to study bacterial community in many hosts including the gut microbiome. The 16S rRNA gene is commonly present in all bacteria consisting of several conserved regions interspersed with non-conserved regions. The primers for PCR amplification are designed from these conserved regions and the non-conserved regions in between them are amplified. Diligent scrutiny of the sequences in the non-conserved region is required as these are specific to each bacterial species or genera. It is these regions that help in identifying taxonomy of the members in the bacterial community. Once all these sequences are obtained there is cleaning of data to ensure all sequences are complete and only rom microbiota and none from the host. After pre-processing, it needs to be sorted into their characteristic bins called Operational Taxonomic Units (OTUs). Sequences with a similarity of 97% are grouped together into an OTU and so on the sequences are clustered into several OTUs (83). The representative sequences are compared with databases containing sequences earlier identified from bacteria and the phylogenetic and taxonomic identity.

Hence, each OTU can be potentially identified as that taxa although the taxonomic identity is not 100% sure. But this approach is very useful to compare microbiome composition under two conditions as it gives information on how different the two communities are devoid of any specific taxonomic identity. A statistical analysis of the proportions of microbial OTUs or organisms can be used to extrapolate the structure of bacterial communities (50).

The main statistical analysis is to find out how diverse a community is. For instance, in the case of analyzing gut microbiome in reference to gastric cancer, the scientist is interested in finding out how far the gut microbiome has deviated from its healthy condition. In other words, he or she needs to know whether any bacterial species or genera has been lost or gained or are there huge differences in the relative abundance of organisms. The ecological measures or indices to understand diversity of macro organisms are used in microbiome studies as well. Several alpha and beta diversity indices can be used; the most suitable to answer one's research question is selected.

There are several limitations to any type of microbial sequence processing methods. From sequencing itself there are huge challenges. All next-generation sequencing techniques have been advanced but are still error-prone than traditional Sanger sequencing methods. The sequencing instrument can be another source of systematic errors (84). Pre- and post-analysis steps needs to be thoroughly checked to ensure only quality reads end up in analysis. Statistics

currently used in biomedical research are traditional, with some improvements that do not handle the intrinsic complexness of a biological system. They are also constrained by the presumption that predictor variables (e.g., gut microbes) are independent (50).

As next-generation sequencing technologies emerge and become more widely available, bioinformatics and computational tools will be more useful for understanding the microbial community composition and capacity in the gut. The downward trend in sequencing costs will allow studies to be conducted on a larger scale, addressing statistical issues. It is imperative to design methods that include patients whose genotypes are well-defined in terms of environmental and dietary factors that affect the gut microbiome. Only by combining gastroenterologists, microbiologists, molecular biologists, computational scientists, bioinformaticians, and statisticians can we successfully address the many bewildering questions we would like to know about the gut microbiome.

2.4.3 ROLE OF R IN BIOINFORMATICS

The bioinformatics pipeline in this project to analyze the gut microbiome composition linked to different types of gastric cancer is designed from tools from three different R packages. They are *Vegan*, *phyloseq* and *microbiome* R packages (85–87). *Vegan*, was originally intended for community ecologists, it is now widely used by microbial ecologists. *Vegan* is not a stand-alone package; it is dependent on various packages in R and must be run in an R statistical environment. *Vegan* also offers tools for diversity analysis and multivariate analysis, as well as other potentially useful functions. Thus, it is tuned for microbiome data analysis, and is widely used in analyzing ecological communities (86). This study uses *vegan* R package to analyze microbial richness and diversity using the alpha diversity functions giving different alpha diversity indicators.

Microbiome package in R is equipped with numerous tools and functions for analyzing microbiome profiling data. It has tools to analyze microbiome data sets and can integrate with other statistical packages. *Microbiome* package provides additional functionality to analyze microbiota composition, bistability, and other diversity indices on microbiome data sets as well as to fit linear models, compare pairs, and analyze associations. The package also contains tools for visualization of results in the form of graphs, plots on ordination axes, heatmaps and other utilities (88).

A number of different statistical packages can be integrated with *phyloseq* to perform statistical hypotheses testing and analysis. *Phyloseq* can analyze taxonomic diversity and perform statistical modelling in conjunction with many R packages, including *DESeq, DESeq2, edgeR*, and *DESeq2.* It also supports data interoperability with other packages and pipelines, such as QIIME2 and Mothur. Additionally, diversity metrics can also be analyzed by *phyloseq* (87,89). A user can perform basic analyses such as beta diversity analysis, alpha diversity analysis, k-table analysis, and differential analysis of microbiome data by importing the data into R. A variety of functions and tools are available to visualize microbiome data with *phyloseq* packages, including bar plots, box plots, density plots, motion charts, ordination plots, and clustering plots (86,89,90).

3. MATERIALS AND METHODS

3. 1 SAMPLE COLLECTION

Samples were taken from patients at Surgical and Meilahti Hospitals in Helsinki Uusimaa Hospital District, Finland. Stool samples from gastric cancer or GIST patients who had no antimicrobial medication during the last six months prior to sample collection and not yet started cancer treatment were collected. There were 23 GIST patients and 29 adenocarcinoma patients in the study. Previously, a study on comparison of microbiota in patients based on location of tumor in gastrointestinal tract (stomach, colon and rectum) was conducted by the same research group (91). Thus, sequencing data 6 GIST and 25 adenocarcinoma patients were also included in this study. A total of 13 stool samples from healthy Finnish adults were taken for controls and their sequencing data used in previous studies (91,92) were included in the present study. Table 3 shows sample details. The collected samples were processed for DNA isolation and next generation sequencing (NGS) using IonTorrent technique.

Table 3: Details of the subjects included in the study. Table taken from (72).

GROUP	NO. OF. CASES	AVERAGE AGE (YEARS)	SEX (M/F)
GASTRIC ADENOCARCINOMA	29	72 ± 11	14/15
DIFFUSE ADENOCARCINOMA	12	69 ± 11	5/7
INTESTINAL ADENOCARCINOMA	15	75 ± 10	8/7
MIXED ADENOCARCINOMA	2	69 ± 11	1/1
GIST	23	67 ± 14	11/12
CONTROLS	13	44 ± 14	3/10

3. 2 16S rRNA GENE SEQUENCING

From stool samples, DNA was extracted, and the quantity and quality of the DNA were determined. 3ng of DNA were used to prepare sequencing libraries. Using primer sets V2, V4, V8, and V3, V6-7, V9, six hypervariable regions of 16S rRNA gene were amplified in two reactions/sample. PCR was followed by end-repair, purification, and ligation of the samples to barcoded sequencing adapters. Samples were diluted to a 10 pM concentration after the libraries were quantified. Libraries were pooled and templates were generated. Afterwards, emulsion PCR was carried out on the libraries. Sequencing was performed on the emulsified libraries. The detailed method of 16S rRNA gene sequencing is explained in (72,92).

3. 3 PIPELINE

From the microbiome data obtained, a bioinformatics pipeline was designed to process and analyze the data. 16S rRNA gene sequenced data was processed after checking the quality. Raw data was filtered using Ion Torrent Suite software (Thermo Fisher Scientific, Waltham, MA, USA). The IonReporter v.5.10 and Metagenomics 16S pipeline w1.1, with default settings, were used to cluster operational taxonomic units (OTUs) from Curated MicroSEQ(R) 16S Reference Library v2013.1 and Curated Greengenes v13.5 databases. The general practice is to cluster sequences with 97% similarity. Here, it is highly advantageous that OTU clustering is computational considering millions of reads obtained from a single sequencing (93).

After the cleaning the data, it is converted and then input to the software packages for analysis in R. The outputs are obtained as many files including sample data (metadata), OTU table (OTU abundance), and taxonomy table (taxonomic ranks) combined together to form the phyloseq object (Figure 2). It was combined using the Bioconductor package, *phyloseq* (87). Phylogenetic sequences based on taxonomic clustering as well as related data types are included in this package, making it very useful. Covariate information, abundance data, and phylogenetics are combined in *phyloseq*. The package has been built following the S4 object-oriented framework of the R language. Data can then be transformed, plotted, and analyzed easily once they have been entered (87,89). The phyloseq object is then used as a template to analyze the microbiome data statistically.

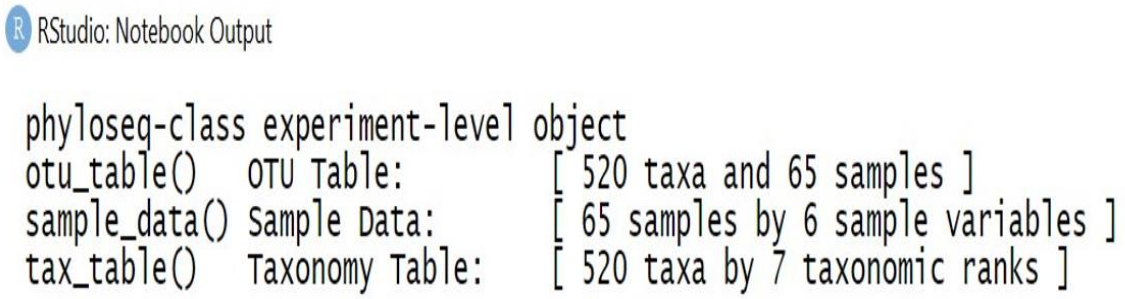


Figure 2: R output showing the structure of the phyloseq object.

There are many other data formats that can be utilized according to the aim of the study, However, in our study the above-mentioned files were needed. The data in the phyloseq object is used as template to analyze and find the differences in composition or diversity of microbiomes of different samples in the study. Structure of phyloseq- class is shown in Figure 3.

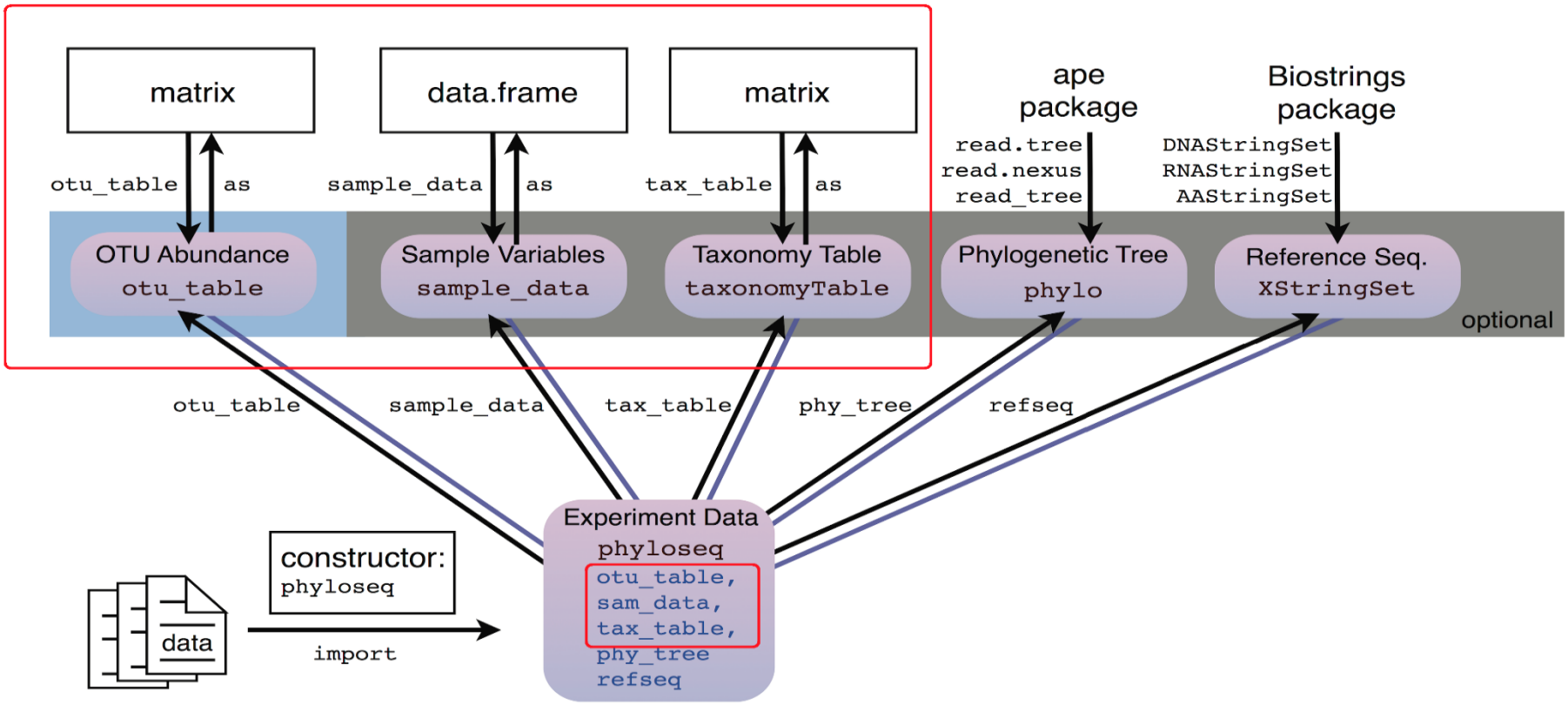


Figure 3: Summary of the structure of the phyloseq-class and its components. The figure is taken from (94) and modified foe this thesis. The components with the highlighted red square are tackled in this thesis. The figure is licensed under CC BY 4. 0.

3. 4 MICROBIOME PACKAGE

Next-generation sequencing studies have revolutionized ways of DNA sequencing and extracting massive amounts of biological information. Although the microbial community consists of bacteria, fungi, viruses and other prokaryotes, most of the studies look at the bacterial population. The 16S rRNA gene profiling technique is used to identify bacterial members in gut microbiome, or otherwise known as taxonomic profiling. Once the DNA sequences of 16S rRNA gene from bacterial members of the gut community is obtained from samples (e.g., stool samples), the raw data is polished, statistically analyzed and visualized using *microbiome* R package, version 1. 14. 0 (85). The *microbiome* R package has various functions to analyze different aspects of the microbial community.

3. 4. 1 ALPHA DIVERSITY

Alpha diversity refers to the species diversity within a habitat or ecosystem. In terms of species richness and evenness. Richness refers to the number of taxonomic groups of the organisms present in the habitat while evenness refers to the distribution of abundances of the groups. Hence alpha diversity metrics usually accounts for the structure of an ecological community in terms of these two components (87). The alpha diversity function ('alpha') estimates various diversity indices for given data. They include Chao1, inverse Simpson index, Gini-Simpson index, Shannon diversity, Fisher's, Camargo evenness, Pielou evenness, Bulla evenness, Evar evenness, absolute dominance, relative dominance, low abundance rarity and many more indices. The various diversity indices can be called out separately with specific functions. For instance, the 'richness' function returns observed richness with given detection thresholds. As the name implies, the dominance index is based on the abundance of species with the highest density. The rarity index measures taxa with low abundance and rarity. A core_abundance function indicates the relative proportion of core species. The 'evenness' function returns a table of various evenness measures such as Camargo, Pielou, Simpson, Evar and Bulla.

While Chao1 and Shannon diversity are the indices used in this study and in most studies to indicate alpha diversity, the alpha diversity function gives various indicators. For instance, another alpha diversity estimator is Simpson index. The Simpson index also integrates richness and evenness. Contrary to the Shannon-Wiener index, it emphasizes common species. Alpha

diversity varies in abundance from 0 to almost 1; the Simpson index is directly proportional to alpha diversity (88). Each researcher may select the index most suitable for their study. It is preferable to select an estimator that is precise and unbiased. In case of microbiota samples, one favours precision which is a simpler property to assess as an exhaustive microbiota sampling is nearly impossible. Since most ecological questions require only relative diversities, most microbiome studies are also answered using relative abundances data (95) and thus exhaustive sampling can be eliminated.

The non-parametric estimators are however, considered most appropriate for studying microbial communities as it is ideal for classes with low abundance in data sets. In other words, in a very diverse community as with animal populations, the occurrence of seeing one species recurrently is rare. Conversely in a microbial community, it is more likely to observe a species repeatedly in a sample, and numerous species will appear multiple times. The Chao1 is non-parametric estimator and uses the above statistical principle to estimate richness by adding a correction factor to the observed number of species (96). It cannot therefore reflect microbial abundance. Shannon-Wiener is a measure of both richness and evenness (97). Rare species are given more weight here, which means it's higher when there are more rare ones.

Wilcoxon-Mann-Whitney test was used to test the significance of differences between the groups and Benjamini-Hochberg FDR was used to correct for multiple testing (72).

3. 4. 2 BETA DIVERSITY

After the initial tasks of OTU clustering and finding microbial richness and diversity, the goal of microbiome studies is to find out how different each sample are from each other. Microbiome studies are more meaningful in understanding whether the communities between groups vary significantly which is accounted by beta diversity. Microbial communities in an environmental context is usually analysed by multivariate statistical methods or models. The environment of the microbial community can be host tissues such as human, animal or plants, or any natural environments such as water bodies or soil. Thus, for analysing the association of any microbial community with its environmental covariates and outcomes, many multivariate statistical tools are available. The microbiome data on which statistical analysis needs to be performed is in the form of OTUs and taxa abundances and using these it is difficult to study the link between microbiome composition and environmental factors. This is because the data

is high dimensional, non-normal and has a phylogenetic structure. Thus, these limitations are highly overcome by multivariate analyses which uses a distance measure method and then performs an analysis of the estimated distances, where the distances between any two microbiome samples are defined as a distance measure (88).

Beta diversity quantifies dissimilarity in community composition between samples. Dissimilarity can be also quantified by distance or divergence. The terms alpha, beta and gamma diversity were introduced by Whittaker in 1960. He defined beta diversity as "the extent of change in community composition, or degree of community differentiation, in relation to a complex-gradient of environment, or a pattern of environments". In terms of diversity, beta diversity is defined as the ratio of gamma (regional) and alpha (local) diversity (98,99) or in other words, it is the number of distinctive compositional units in a region (100). For beta diversity analysis, the PERMANOVA is one of the most commonly applied nonparametric distance-based methods to identify the relationships between microbiome composition and covariates of interest in microbiome data. Permutation and distance matrices are used in PERMANOVA, which is accessible from the R package *vegan* by using the function *adonis* (86). The test statistics include P-value of the permutation test that checks the statistically significant difference in beta diversity between communities. R^2 is a variable that shows the measure of the variance explained by a grouping factor (86). Beta diversity results were visualized using unsupervised principal coordinates analysis (PCoA) using Bray–Curtis dissimilarity index (87).

As described, PERMANOVA is a distance-based statistical analysis. PERMANOVA results of microbiome data are visualized by principal coordinate analysis (PCoA). The reason for this is that PCoA is based on the distance between data points rather than principal component analysis (PCA). As a first step, the PCoA projects the distances into Euclidean space in a greater number of dimensions; for n data points, there are n-1 dimensions. In PCoA, the first point is placed at the origin, the second along the first axis, the third added along the second axis, and so on, until all the points are added. Users appreciate the understanding of beta diversity results visually for which PCoA is applicable. By plotting communities in different dimensions and superimposing the axes similar communities cluster together while distinct communities are separated. PERMANOVA is used to test this more formally (statistically). This helps in distinguishing visually whether the microbial communities in the gut are similar to each other in healthy individuals or similar to each other in diseased patients. The other

important aspect would be to visualize how different the gut microbial communities of diseased individuals are from healthy people. The concept of diversity indices is captured in Figure 4

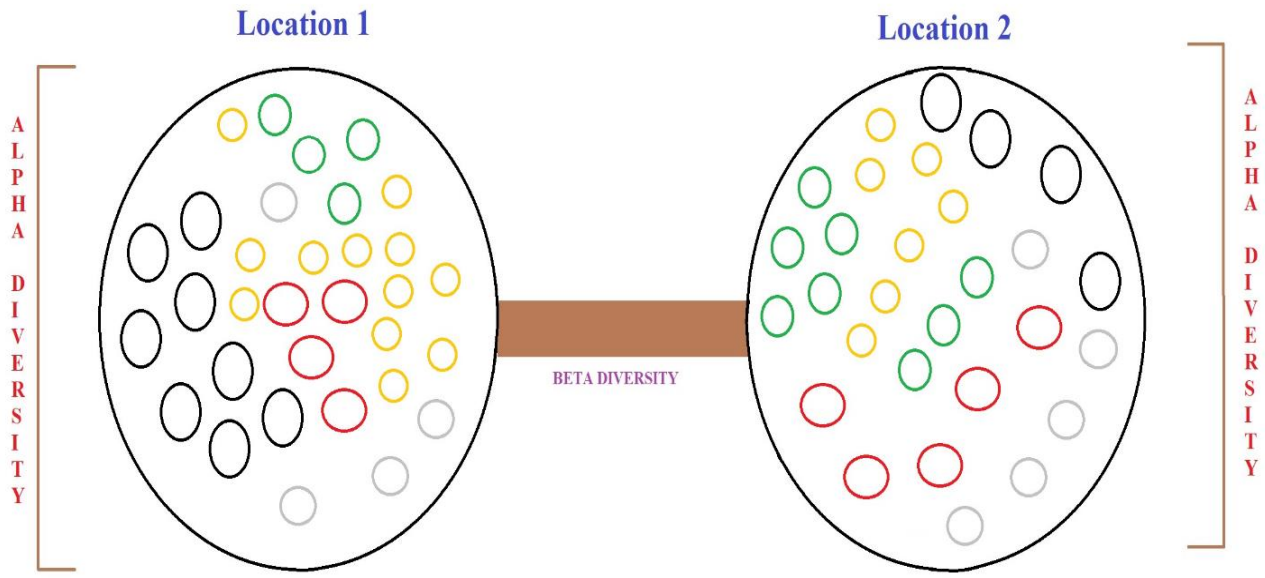


Figure 4: Alpha diversity provides the evenness and richness of units within a habitat (here Location 1 and Location 2) while, beta diversity measures diversity between habitats (here diversity between Location 1 and Location 2).

3. 4. 3 DIFFERENTIAL ABUNDANCE ANALYSIS

As explained in the background (or introduction), numerous diseases are associated with changes that occur in the microbiome or microbial environment. Consequently, researchers are interested in understanding changes in the microbial composition under various conditions. ALDEx2 was used in this study to identify differentially abundant taxa that distinguished healthy gut microbial communities from microbial communities in the stool samples of adenocarcinoma and GIST patients. A taxon in a sample is defined to be differentially abundant between two environments if its absolute mean abundance differs between the environments. Parameter selection for statistical analysis determines whether the differential abundance analysis is between absolute or relative abundances of the taxa. The most important criterion for differential abundance analysis is the sample size to be taken from the ecosystem. It takes into account the bias introduced by differences in sampling fractions between samples. Identification of taxa that differ in mean absolute abundance per unit volume between two or

more habitats answer intriguing questions as to the cause of disease, as in this case - gastric cancer. There are several methods for differential abundance analysis such as *ALDEx2, edgeR, DESeq2* etc. Methods such as *edgeR* and *DESeq2* tend to be sensible and suitable for gene expression data but ineffective with microbiome data. The reason behind this is they use normalization methods that assume an extremely small percentage of taxa are differentially abundant which is not always true for microbiome data. Therefore, these methods provide inherently biased test statistics under the null hypothesis. Hence, *ALDEx2* is preferred for microbiome data as it is designed as a compositional data analysis tool for high-throughput sequencing data like 16S rRNA gene sequencing (101).

The *ALDEx2* R package has been shown to be a simple and robust tool when used with high-throughput sequencing datasets containing per-feature counts. It uses Bayesian method to deduce technical and statistical error. The approach of *ALDEx2* was found suitable for human microbiome 16S rRNA gene data as it correctly identifies differential abundance of OTUs. *ALDEx2's* design decreases the number of false positives that may arise when datasets consist of many features in a few samples (102). It benefits from the clr transformation for relative abundances to dismiss compositionality bias and delivers empirical p-values with Benjamini–Hochberg FDR correction. In other words, since it uses the log ratio transformation methods to transform microbiome data removing the constraints caused by standard multivariate techniques for analysis. One of the problems of microbiome data is it is highly dimensional and multiple comparisons are used. For microbiome data, p-value adjustments are necessary if multiple groups or variables are compared to limit false positives. ALDEx2 tackles this problem of multiple comparisons by correcting the FDR, i.e., the expected proportion of type I errors or the proportion of false positives arising from all rejected null hypotheses. FDR is 5% if five out of every 100 hypothesis tests are false discoveries. The “Benjamini-Hochberg (BH) adjusted P-values” favored over P-values in analyzing microbiome data. The test is considered to be significant when the adjusted P-value is smaller than the chosen FDR (103). A probabilistic sampling procedure is used to estimate the standardized effect size and p-value (73). With *ALDEx2*, you can analyze virtually any data produced by high-throughput sequencing and compare many different experimental designs, compared to previous statistical tools. The statistical analyses comprise of two-sample and paired t-test, ANOVA, and non-parametric tests, like Kruskal-Wallis test, Wilcoxon test, Welch’s t-test. A snapshot of the created pipeline is given in Figure 5.

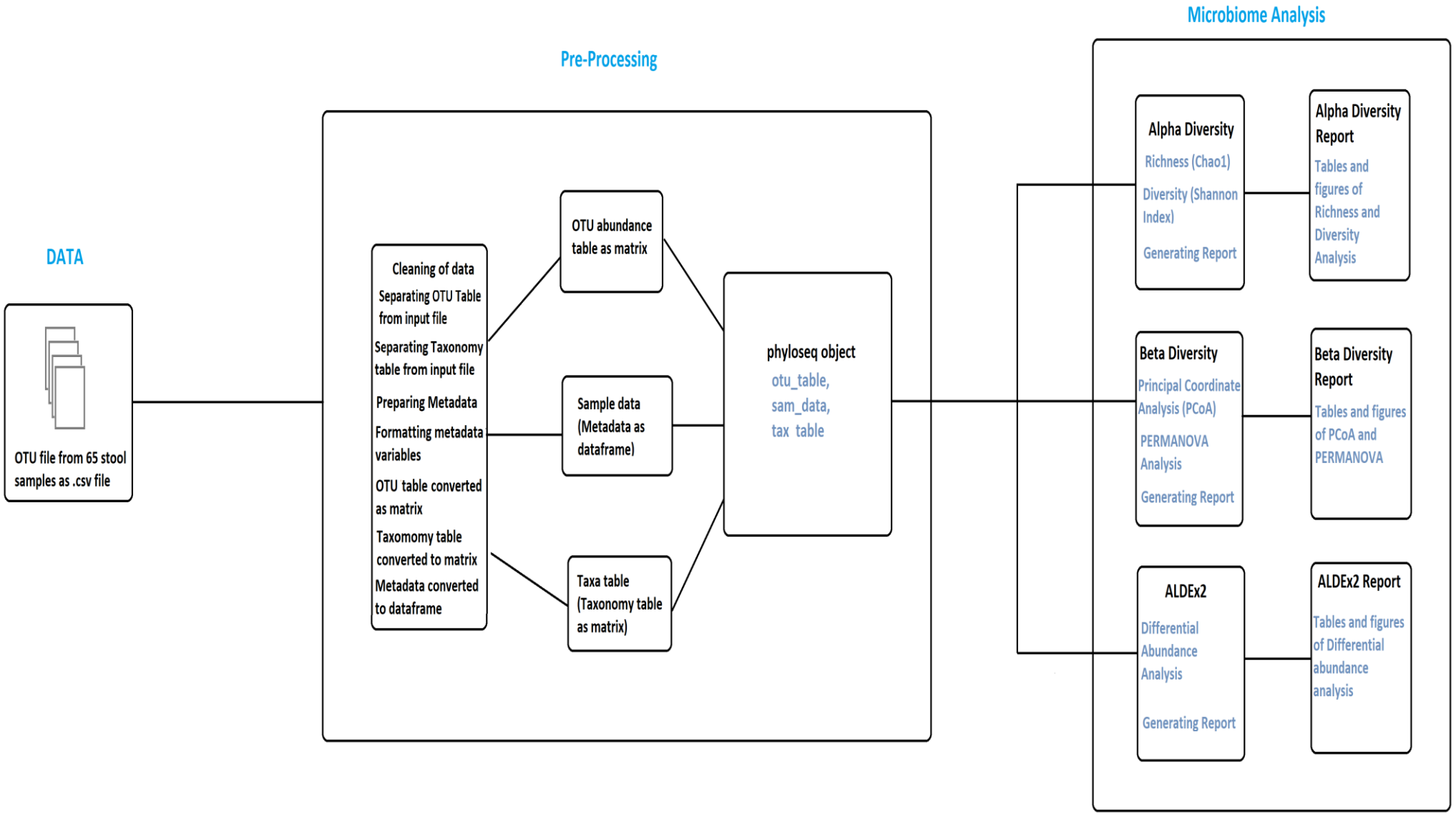


Figure 5: Snapshot of the pipeline created for gut microbiome analysis.

4. RESULTS

4. 1 MICROBIOTA DIVERSITY

4. 1. 1 ALPHA DIVERSITY

In this study, Chao1 was used to quantify species richness and Shannon diversity was used to find species diversity. Microbiota richness (Chao1) (Figure 6) was lower in stool samples for gastric adenocarcinoma and gastric GIST patients compared to controls. Significant differences were observed between adenocarcinoma patients and control samples as well as between GIST patients and controls, while no significant changes between adenocarcinoma and GIST patients.

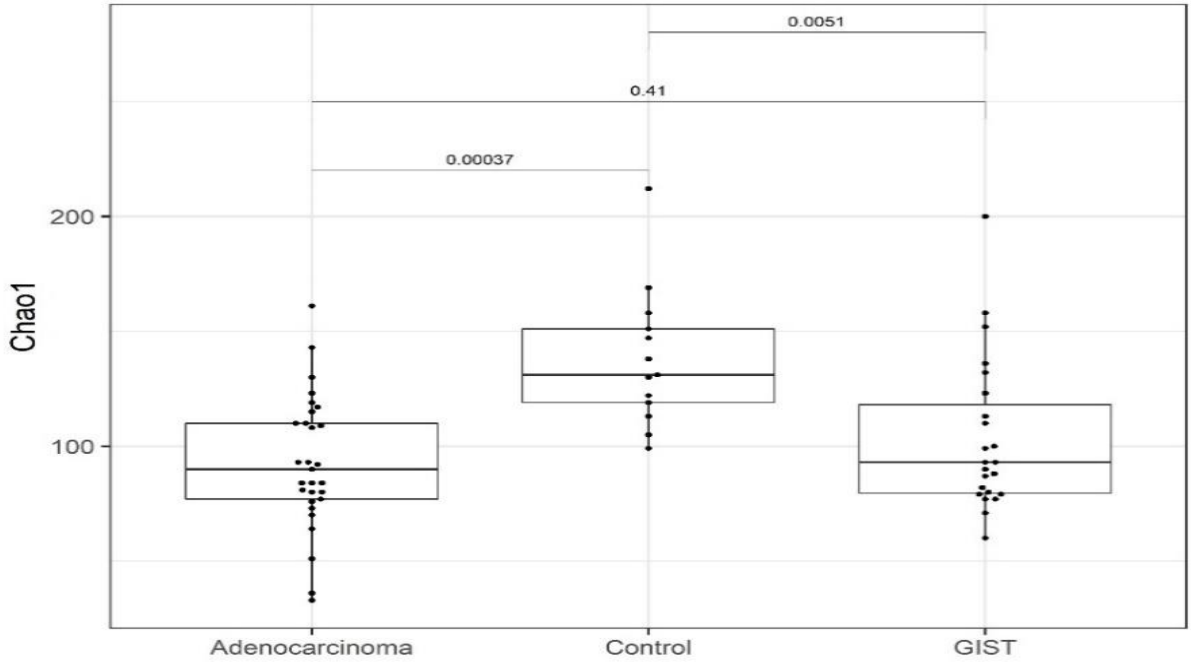


Figure 6: Microbiota richness in the gut microbiome of patients with gastric cancer types compared to control. P-values calculated by Wilcoxon test. Figure taken from (72). The figure is licensed under CC BY 4. 0.

The microbial diversity indicated by Shannon index (Figure 7) was also lower for adenocarcinoma and GIST patients when comparing each group to control. However, the

changes were not significant. The diversity changes between cancer groups were also non-significant.

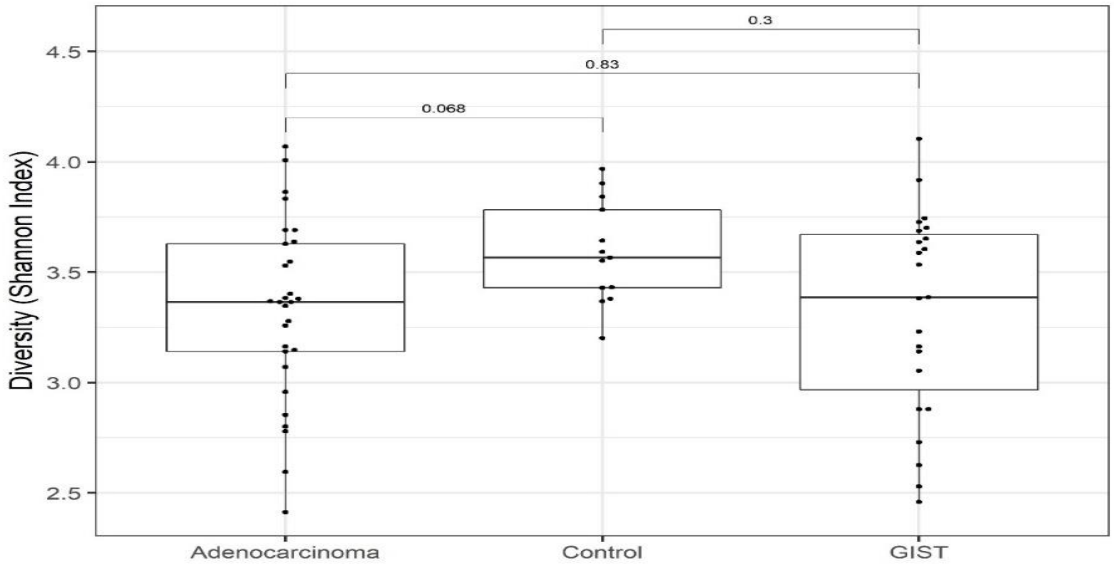


Figure 7: Microbiota species diversity in gut microbiome of patients with gastric cancer types compared to control. P-values calculated by Wilcoxon test. Figure taken from (72). The figure is licensed under CC BY 4. 0.

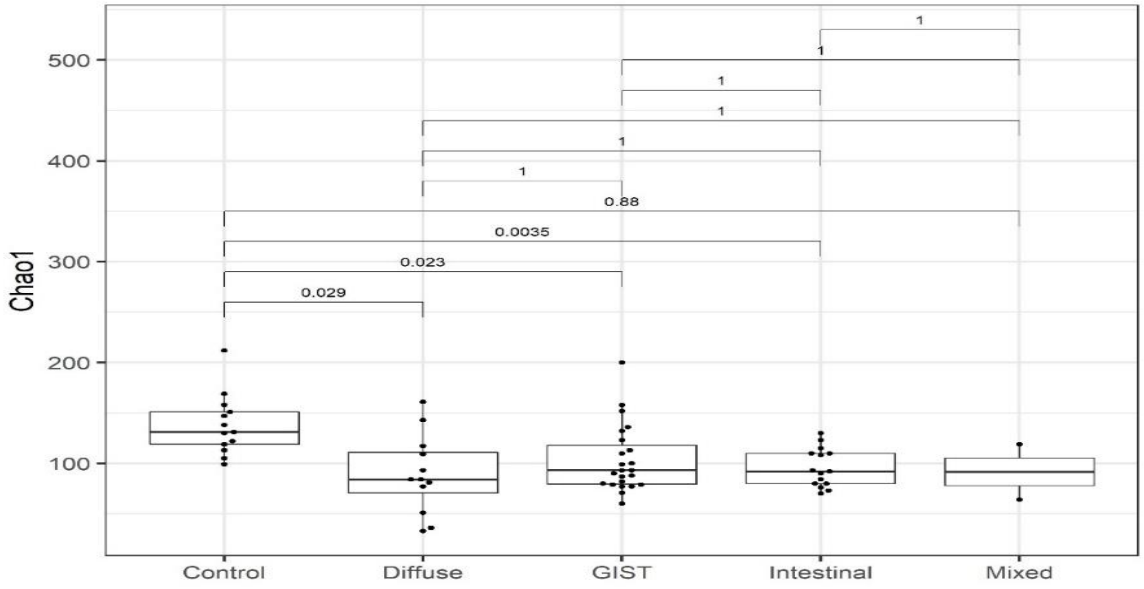


Figure 8: Microbiota richness in the gut microbiome of patients with carcinoma subtypes compared to control. P-values calculated by Wilcoxon test. Figure taken from (72). The figure is licensed under CC BY 4. 0.

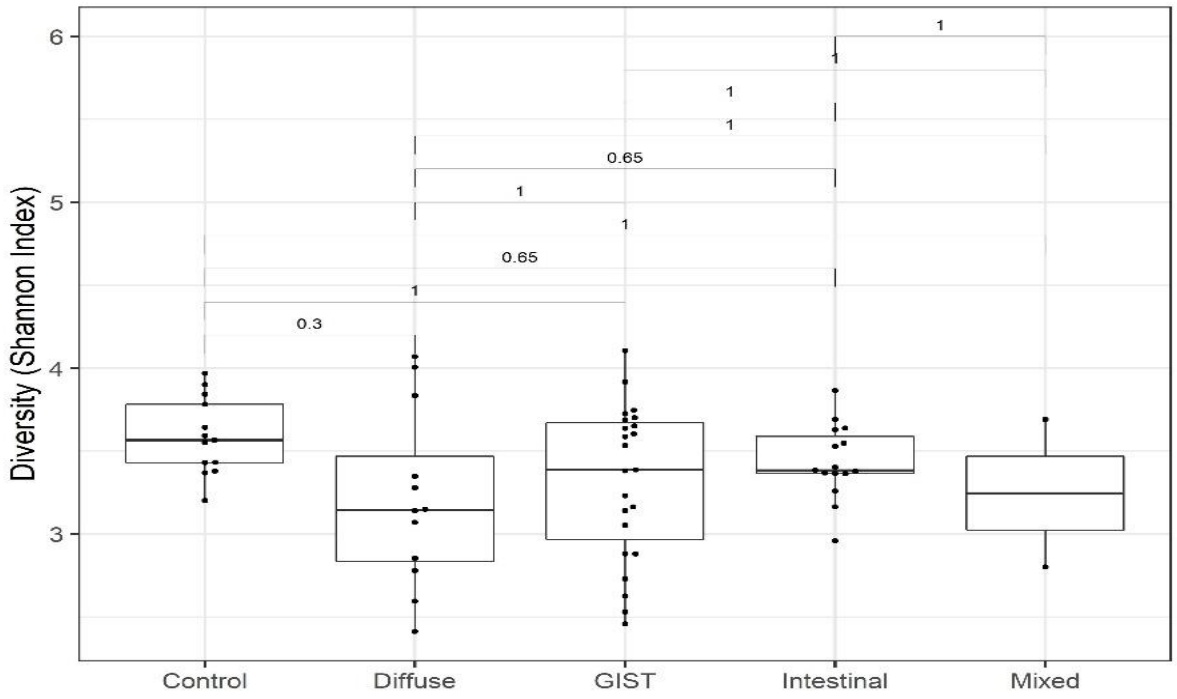


Figure 9: Microbiota diversity in the gut microbiome of patients with carcinoma subtypes compared to control. P-values calculated by Wilcoxon test. Figure taken from (72). The figure is licensed under CC BY 4. 0.

Microbiota richness and diversity was studied for different subgroups (diffuse, intestinal, GIST and mixed) of diffuse adenocarcinoma patients. The richness was significantly lower for all the subgroups compared to control (Figure 8). Although diversity was lower for cancer subgroups compared to control, they were not significant (Figure 9). The richness and diversity did not show any significant differences between subgroups.

4. 1. 2 BETA DIVERSITY

The PERMANOVA analysis between controls and GC types showed that significant (p=0. 03) differences between control and GIST groups (Table 4). These observations are visualized in the PCoA graph wherein the GIST and adenocarcinoma samples deviating from the control group cluster (Figure 10).

Table 4: Results of PERMANOVA analysis showing significant changes in gut microbiome of cancer patients compared to control.

Comparison	R2	Adjusted p-value
Adenocarcinoma *vs* Control	0. 049	0. 03
Adenocarcinoma *vs* GIST	0. 018	0. 55
GIST *vs* Control	0. 046	0. 03

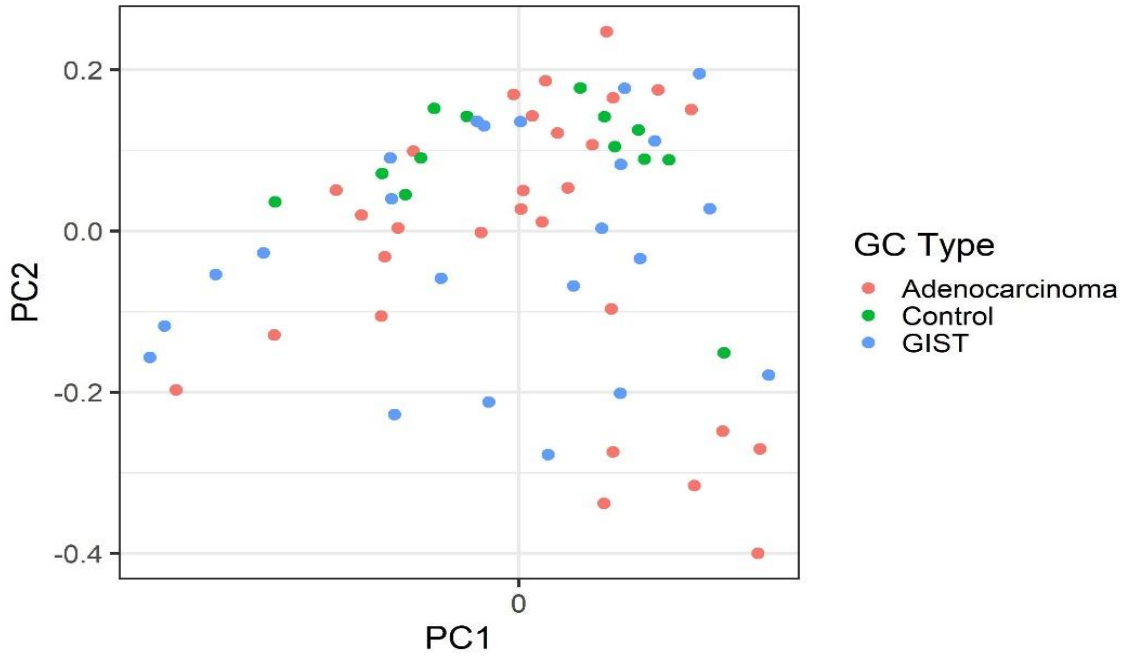
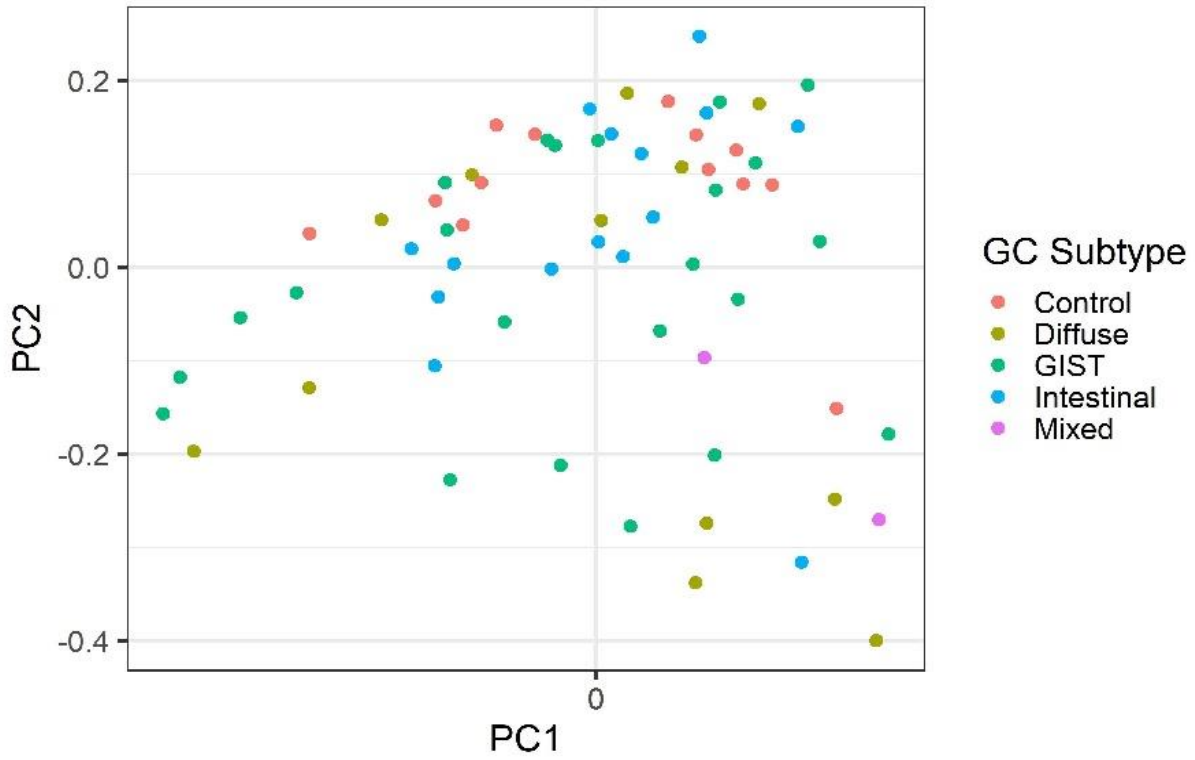


Figure 10: Principal Coordinate Analysis (PCoA) based on Bray-Curtis on gut microbiome in patients with gastric cancer types compared to control.

The differences in microbiota composition between controls and GC subtypes were not significant (Table 5). Between GC subtypes, the variation in microbiota composition was checked at the genus level using Bray-Curtis dissimilarity index, but was not significant. This is clearly depicted in the PCoA graph (Figure 11).

Table 5: Results of PERMANOVA analysis comparing in gut microbiome of cancer subtypes. Table taken from (72).

Comparison	R2	Adjusted p-value
Control *vs* Diffuse adenocarcinoma	0. 08	0. 05
Control *vs* GIST	0. 05	0. 05
Control vs Intestinal adenocarcinoma	0. 07	0. 05
Control *vs* Mixed adenocarcinoma	0. 13	0. 05
Diffuse adenocarcinoma *vs* GIST	0. 02	0. 89
Diffuse *vs* Intestinal adenocarcinoma	0. 04	0. 58
Diffuse adenocarcinoma *vs* Mixed	0. 07	0. 66
GIST *vs* Intestinal adenocarcinoma	0. 03	0. 32
GIST *vs* Mixed adenocarcinoma	0. 04	0. 54
Intestinal vs Mixed adenocarcinoma	0. 08	0. 20



.

Figure 11: Principal Coordinate Analysis (PCoA) based on Bray-Curtis on gut microbiome of patients with carcinoma subtypes compared to control. Figure taken from (72). The figure is licensed under CC BY 4. 0.

4. 1. 3 DIFFERENTIAL ABUNDANCE ANALYSIS

4. 1. 3. a. DIFFERENTIAL ABUNDANCE OF TAXA AT FAMILY LEVEL IN ADENOCARCINOMA, GIST AND CONTROL

ALDEx2 differential abundance analysis was performed to find bacterial taxa that significantly differentiated groups in terms of relative abundance in pair-wise comparisons (Table 6). Bacteria belonging to Enterobacteriaceae was highly abundant in adenocarcinoma and GIST patients compared to control, while the Lactobacillaceae bacteria was relatively lower in the both the cancer groups compared to controls. In addition, Oscillospiraceae, Bifidobacteriaceae and Eubacteriaceae showed significant lower abundance in adenocarcinoma patients compared to control. In GIST patients, other than the above-mentioned taxa no other bacterial families were significantly detected. Visual representation was done by plotting the significant taxa against their centered log-ratio (clr) transformation (Figure 12).

Table 6: Results of ALDEx2 differential abundance analysis showing most abundant family in gastric cancer types.

Family	P-value	Effect	Groups compared
Lactobacillaceae	0. 01	0. 64	Adenocarcinoma *vs* Control
Enterobacteriaceae	0. 01	-1. 22	Adenocarcinoma *vs* Control
Oscillospiraceae	0. 01	0. 64	Adenocarcinoma *vs* Control
Bifidobacteriaceae	0. 03	0. 85	Adenocarcinoma *vs* Control
Eubacteriaceae	0. 04	0. 71	Adenocarcinoma *vs* Control
Lactobacillaceae	0. 03	-0. 75	Control *vs* GIST
Enterobacteriaceae	0. 03	1. 01	Control *vs* GIST
Oscillospiraceae	0. 05	-0. 47	Control *vs* GIST

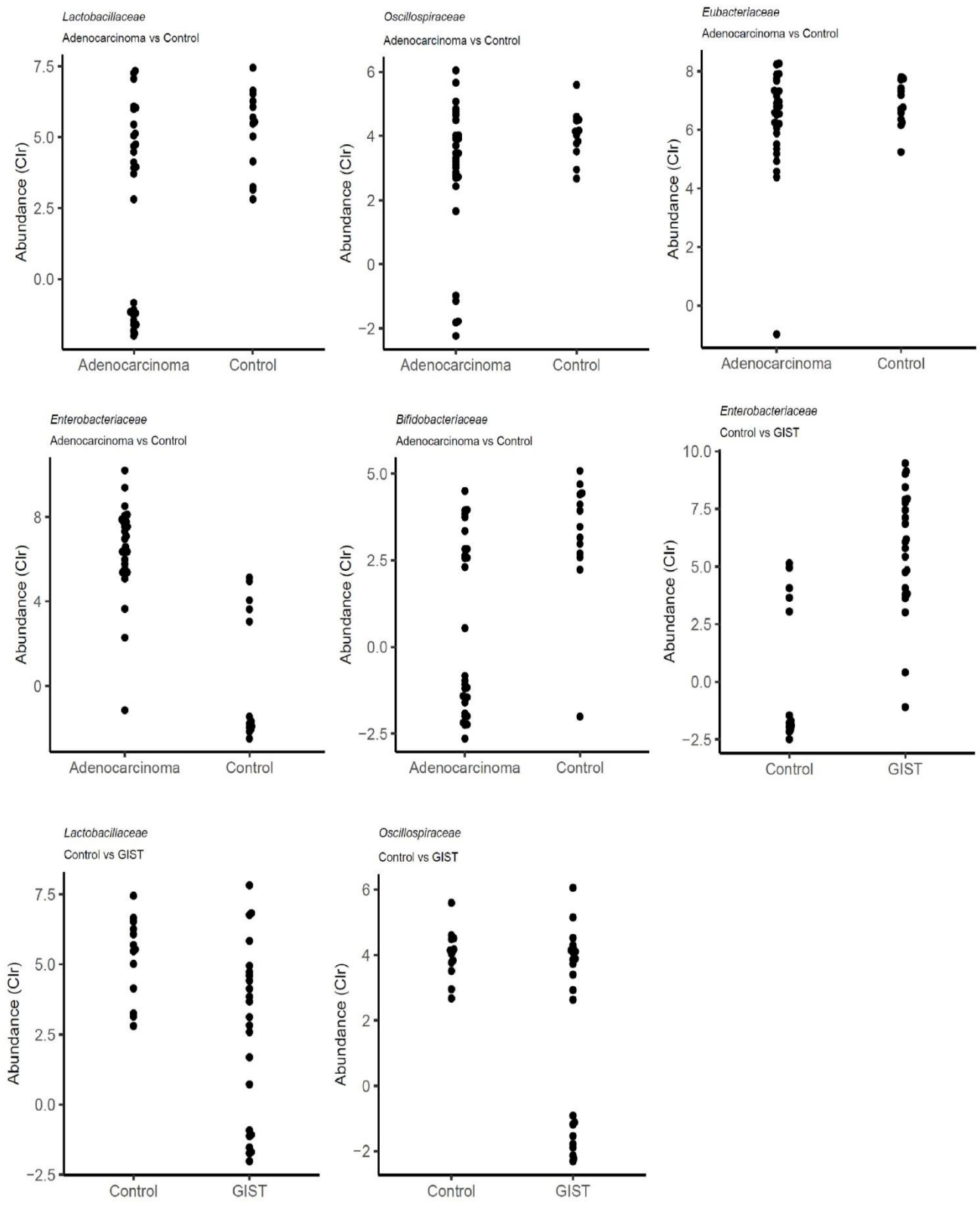


Figure 12: Bacterial taxa **(**Family level) with significant differences in pairwise comparison between controls, adenocarcinoma and GIST.

4. 1. 3. b DIFFERENTIAL ABUNDANCE BETWEEN ADENOCARCINOMA SUBGROUPS

Microbiota composition between controls and adenocarcinoma subgroups (diffuse, intestinal and mixed) and between groups were compared. It showed taxa with significant differential abundance in various pair-wise groups (Table 7). Bacterial family Enterobacteriaceae was significantly higher in both the adenocarcinoma groups – mixed, diffuse and intestinal compared to controls. Similar trend was observed with significant higher abundance of Lactobacillaceae, Oscillospiraceae and Bifidobacteriaceae compared to controls (Figure 13).

Table 7: Results of ALDEx2 differential abundance analysis showing most abundant family in gastric cancer subtypes.

Family	P-value	Effect	Groups compared
Enterobacteriaceae	0. 02	1. 24	Control *vs* Diffuse adenocarcinoma
Bifidobacteriaceae	0. 05	-1. 13	Control *vs* Diffuse adenocarcinoma
Lactobacillaceae	0. 03	-0. 76	Control *vs* GIST
Enterobacteriaceae	0. 03	1. 02	Control *vs* GIST
Oscillospiraceae	0. 05	-0. 46	Control *vs* GIST
Enterobacteriaceae	0. 04	1. 15	Control *vs* Intestinal adenocarcinoma
Enterobacteriaceae	0. 05	1. 95	Control vs Mixed

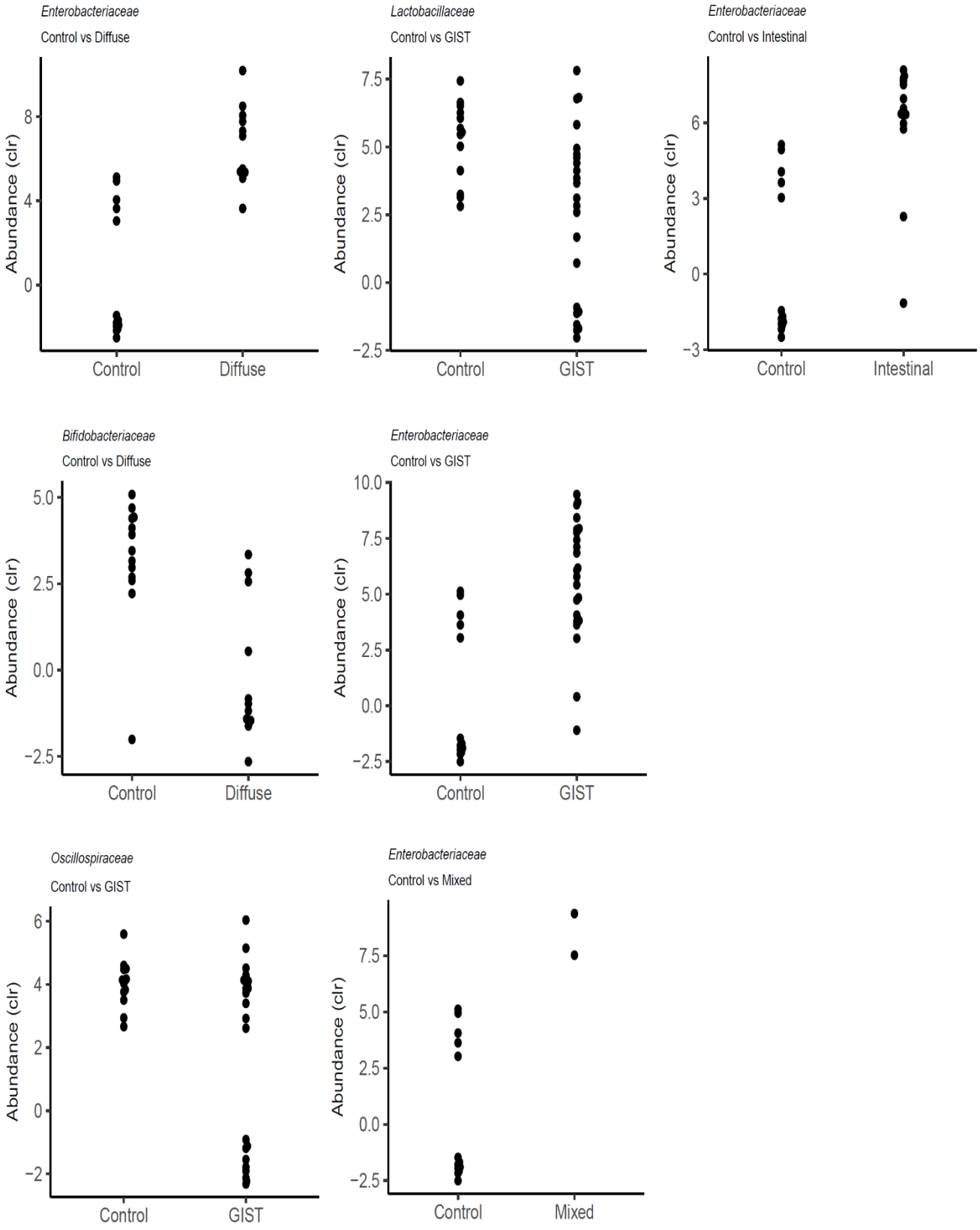


Figure 13: Bacterial taxa **(**Family level) with significant differences in pairwise comparison between controls and gastric subgroups.

5. DISCUSSION

5.1 BIOINFORMATICS IN MICROBIOME ANALYSIS

Human genome project was a huge leap in various disciplines related to health (104). However, there had been certain grey areas where we could not understand the links between health or behaviour and certain gene functions. These knowledge gaps are now being filled in by the advances in human microbiome research. Microbiome studies has advanced our knowledge on human health and behaviour (17).

This advancement in microbiome studies is owed to the rapid progress in DNA sequencing techniques. The commencement of next generation sequencing techniques has helped to capture a snapshot of the microbiome in any target tissue. The last two decades had been bombarded with numerous studies in human microbiome – related to different tissues mainly skin, mouth and gut. Sequencing the microbiomes in these tissues yielded more than half-a-million reads and huge data were derived from these sequences. The primary objective of these studies was to understand the structure and composition of the microbial community in a healthy person and how it differed in a patient with certain ailment. The crucial step is to identify the bacterial or fungal species associated with each tissue and what species are lost or gained when the tissue deviated from homeostasis (17). For instance, a patient with gastric cancer has gut microbiota different from a healthy person and this condition is referred to as dysbiosis. Hence, from sequencing studies, researchers are very much interested in identifying the taxonomy of microbes residing on human tissues.

The bacterial community is of particular interest to scientists since there is enough scientific literature available onto which new studies may be built up. In addition, the technique of sequencing microbiota using 16S rRNA gene is very convenient. However, from the commencement of studies using 16S rRNA targeted gene sequencing, it has become very clear that huge amounts of data are generated from microbiome sequencing.

Microbiome data collected is mainly to characterize the association between microbiome characteristics and biological, genetic, and clinical conditions; and to determine factors associated with microbiome composition based on biological and environmental

influences. These studies aim to understand how genetic and environmental factors influence our microbiome (88)

This data needs to be organized and statistically analysed to bring out meaningful interpretation, which brings in the necessity to convert biological data to computerized data, in order to operate on the data and finally retrieve a biologically meaningful. Therefore, there is a need for a pipeline to guide the researcher on how to proceed once the sequences are obtained.

5.2 DATA INTERPRETATION FROM THE ANALYSIS

The main aim of this project was to design a pipeline guiding on the steps to biologically interpret the gut microbiome sequences obtained from patients with gastric cancer. Samples were collected from patients with two types of gastric cancer – adenocarcinoma and GIST, which included five different subtypes of adenocarcinoma as well.

One of the important parameters determining the structure and composition of a microbial community is species richness and diversity of its microbial members. The alpha diversity function gives several measures of microbial species richness and diversity. Chao1 index takes into account microbiota richness. It considers the total unique species, the number of singleton taxa, and doubleton taxa. A healthy gut is characterized with higher microbial gene richness which is translated into higher microbial richness (105). Richness alone does not give an overall account of the gut microbiome; hence the diversity is also taken into consideration.

The Shannon diversity index is relevant in this context since it correlates both species richness and evenness evaluating the species diversity. As the number of rare species rises, it has a greater weighting, which means it is higher. A higher value indicates greater alpha diversity. A high taxa diversity characterizes healthy gut microbiome (106), although later studies have proven that it can be a biased indicator (107). However, species diversity still remains a valuable indicator of healthy microbiome versus diseased microbiome. In our study also, the microbiota richness and diversity were significantly lower in patients with gastric cancer compared to healthy individuals. Lower alpha diversity has been observed in gut microbiome of breast cancer patients with human epidermal growth factor receptor 2 (HER2+) compared to HER2- (109). Not only with cancer, lower microbial richness and diversity has

been observed in other unhealthy conditions such as malnutrition, obesity or related to diseases such as type 2 diabetes, ADHD (Attention Deficit Hyperactivity Disorder) (105).

The results in this study shows a decrease in richness and diversity in cancer patients compared to controls. The next logical step in understanding the structure and composition of microbial community in the gut is to find out whether this is due to loss or gain of bacterial species or vast difference in the abundances of certain microbial taxa when comparing with the gut microbiome of healthy individuals. This is relevant not only in gut microbiome studies but in most human microbiome studies. It is essential to find how much the microbial community composition has deviated from the healthy microbiome. These differences were statistically analysed using PERMANOVA (Permutational Multivariate Analysis of Variance) test. As in the present study, we find that the microbiome samples of adenocarcinoma and GIST patients significantly differed from control samples. Analyses of beta diversity also revealed that all cancer subgroups had significantly different microbiota compositions when compared to controls. A subsequent decrease in the richness of microbiota has been observed in early stages of gastric cancer with significant differences in GC samples compared to control (108,109) .

PERMANOVA has been used in other studies related to gut microbiota such as to find the difference in the gut microbiota composition in omnivore versus vegans (110), to test for microbial variations among different populations (111), linking role of gut microbiota to food allergies namely egg (112) and cow's milk (113), to determine the influence of ethnicity on gut microbiome of individuals sharing a geographical location (21) or identifying genetic variations from gut microbiome (114,115).

The presence of biologically meaningful results does not always require increased sequencing depth. Modern sequencing methods enable researchers to discern differences between samples even at relatively low sequence coverage when they choose diversity measurements appropriate for their study (12).

It is also important to identify microbial members that maybe relatively higher or lower in abundance in dysbiosis conditions. This is vital because they may be used as biomarkers for preventive or therapeutic approaches to identify gastric cancer at earlier stages. The differential abundance analysis in this study was analyzed using the ALDEx2 package. In the present study,

with differential abundance analysis significant bacterial taxa at the family as well as genus levels could be identified in GC types and subtypes compared to the controls.

5.3 LIMITATIONS AND FUTURE PROSPECTS

Even though the pipeline was able to find the microbiome association in gastric cancer patients which aligns with previous studies, it is important to understand that these analyses provide the association of the microbiome; not the cause of the disease. The fact that the analysis is done on the taxonomic level and not on a functional level is a limitation of this pipeline. Although 16S rRNA gene sequencing is widely used due to recent advances and benefits, aberrations are occurring at various stages of molecular experimentation, including error-prone PCRs, and biases introduced during data analysis, such as OTU clustering, reference databases, and specific software implementations (116,117). Taxonomic classification and microbiome analysis may be seriously affected by these methodological differences (118).

Apart from standard microbiome analysis methods (alpha and beta diversity), I have used *ALDEx2* for differential abundance analysis instead of popular methods like *DESeq2*, baySeq, *ANCOM* and *edgeR*. The advantages of *ALDEx2* method were discussed in the material and methods part, but it also has some limitations. The first disadvantage is the long runtime of ALDEx2 compared to edgeR and other methods because non-parametric analyses are replicated across multiple Monte Carlo instances. The second issue is that ALDEx2 lacks a recorded generalization to mixed models (119).

Based on the study design, one needs to make minor adjustments to the pipeline. According to the source of the OTU clustering, the pre-processing stage must be modified. There is a wide variety of formats in which OTU tables are generated. The OTU table used here was created by IonReporter, which generated an OTU table for each sample, and then combined the OTU tables in the pre-processing stage. The microbiome analysis part is straightforward, and one can adjust the variables and plot settings according to their needs.

As mentioned earlier, microbiome analysis is evolving, and as long as the standard analysis approach is utilized, this pipeline can be used. This pipeline is not limited to the analysis of gut microbes. As a pipeline uses standard microbiome analysis methods, it can

analyze the microbiome from any source after customization (such as stratification of samples into subsets, or updates in the study covariates). Access to the pipeline is available upon request.

There are other sources of microbiome aside from gut, but further studies and benchmarking would need to be conducted in order to evaluate the efficiency of this approach in other microbiome studies. Further research is needed in order to find ways to extend the functionality of the pipeline in terms of pathways analysis, time series analysis, and community structure of microbiota

6. CONCLUSION

This study creates a pipeline to identify the structure and composition of gut microbiota data related to gastric cancer. The pipeline has been designed using statistical tools from vegan, phyloseq and microbiome R packages. The created pipeline merges the relevant sections from each of these above packages for the convenience of the user or researcher. Allowing each user can select the statistical outputs and visualization plots most appropriate for their analysis.

The pipeline streamlines the data analysis of a data set from pre-processing until the diversity analysis stage. The pipeline is divided into three main sections of analysis: alpha diversity, beta diversity, and composition analysis. For alpha diversity and beta diversity analysis, a variety of analyses are possible, allowing several estimators to be computed by the user. For the composition analysis, the differential abundance analysis is described in the thesis explaining how to find prominent bacterial taxa at the family and genus level.

The pipeline described in this thesis will help upcoming researchers in the future as a simple tool for their analysis, allowing them to further their research smoothly.

7. REFERENCES

1. Salvucci E. Microbiome, holobiont and the net of life. Vol. 42, Critical Reviews in Microbiology. Taylor and Francis Ltd; 2016. p. 485–94.

2. Sender R, Fuchs S, Milo R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. PLOS Biology. 2016 Aug 19;14(8): e1002533.

3. Grice EA, Segre JA. The human microbiome: Our second genome. Vol. 13, Annual Review of Genomics and Human Genetics. 2012. p. 151–70.

4. Gauthier J, Vincent AT, Charette SJ, Derome N. A brief history of bioinformatics. Briefings in Bioinformatics. 2019 Nov 27;20(6):1981–96.

5. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, et al. Diversity of the human intestinal microbial flora. Science (New York, NY). 2005 Jun 10;308(5728):1635.

6. Arnold JW, Roach J, Azcarate-Peril MA. Emerging technologies for gut microbiome research. Trends in microbiology. 2016 Nov 1;24(11):887.

7. Galloway-Peña J, Hanson B. tools for analysis of the microbiome. Digestive diseases and sciences. 2020 Mar 1;65(3):674–85.

8. LEDERBERG J, MCCRAY AT. `Ome Sweet `Omics--A genealogical treasury of words. The Scientist. 2001 Apr 2;15(7):8–8.

9. Fierer N, Ferrenberg S, Flores GE, González A, Kueneman J, Legg T, et al. From Animalcules to an Ecosystem: Application of Ecological Concepts to the Human Microbiome. 101146/annurev-ecolsys-110411-160307. 2012 Nov 5; 43:137–55.

10. Kilian M, Chapple ILC, Hannig M, Marsh PD, Meuric V, Pedersen AML, et al. The oral microbiome – an update for oral healthcare professionals. British Dental Journal 2016 221:10. 2016 Nov 18;221(10):657–66.

11. Bäckhed F, Roswall J, Peng Y, Feng Q, Jia H, Kovatcheva-Datchary P, et al. Dynamics and stabilization of the human gut microbiome during the first year of life. Cell Host and Microbe. 2015 May 13;17(5):690–703.

12. Ursell LK, Metcalf JL, Parfrey LW, Knight R. Defining the human microbiome. Nutrition Reviews. 2012 Aug;70(SUPPL. 1).

13. Lax S, Smith DP, Hampton-Marcell J, Owens SM, Handley KM, Scott NM, et al. Longitudinal analysis of microbial interaction between humans and the indoor environment. Science (New York, NY). 2014 Aug 29;345(6200):1048.

14. Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch S v., Knight R. Current understanding of the human microbiome. Nature Medicine. 2018 Apr 10;24(4):392–400.

15. Caporaso JG, Lauber CL, Costello EK, Berg-Lyons D, Gonzalez A, Stombaugh J, et al. Moving pictures of the human microbiome. Genome Biology 2011 12:5. 2011 May 30;12(5):1–8.

16. Kort R, Caspers M, van de Graaf A, van Egmond W, Keijser B, Roeselers G. Shaping the oral microbiota through intimate kissing. Microbiome 2014 2:1. 2014 Nov 17;2(1):1–8.

17. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The Human Microbiome Project. Vol. 449, Nature. Nature Publishing Group; 2007. p. 804–10.

18. Risely A. Applying the core microbiome to understand host–microbe systems. Vol. 89, Journal of Animal Ecology. Blackwell Publishing Ltd; 2020. p. 1549–58.

19. Chowdhury S, Fong SS. Computational modeling of the human microbiome. microorganisms 2020, Vol 8, Page 197. 2020 Jan 31;8(2):197.

20. Gupta VK, Paul S, Dutta C. Geography, ethnicity or subsistence-specific variations in human microbiome composition and diversity. Frontiers in Microbiology. 2017 Jun 23;0(JUN):1162.

21. Dwiyanto J, Hussain MH, Reidpath D, Ong KS, Qasim A, Lee SWH, et al. Ethnicity influences the gut microbiota of individuals sharing a geographical location: A cross-sectional study from a middle-income country. Scientific Reports. 2021 Dec 1;11(1).

22. Ahn J, Hayes RB. Environmental influences on the human microbiome and implications for noncommunicable disease.101146/annurev-publhealth-012420-105020. 2021 Apr 2; 42:277–92.

23. Bunyavanich S, Berin MC. Food allergy and the microbiome: current understandings and future directions. The Journal of allergy and clinical immunology. 2019 Dec 1;144(6):1468.

24. Abrahamsson TR, Wu RY, Jenmalm MC. Gut microbiota and allergy: The importance of the pregnancy period. Pediatric Research 2015 77:1. 2014 Oct 13;77(1):214–9.

25. Ramirez J, Guarner F, Bustos Fernandez L, Maruy A, Sdepanian VL, Cohen H. Antibiotics as major disruptors of gut microbiota. frontiers in cellular and infection microbiology. 2020 Nov 24; 10:731.

26. Bailey MJ, Naik NN, Wild LE, Patterson WB, Alderete TL. Exposure to air pollutants and the gut microbiota: A potential link between exposure, obesity, and type 2 diabetes. Gut Microbes. 2020 Sep 2;11(5):1188.

27. Fouladi F, Bailey MJ, Patterson WB, Sioda M, Blakley IC, Fodor AA, et al. Air pollution exposure is associated with the gut microbiome as revealed by shotgun metagenomic sequencing. Environ Int. 2020; 138:105604.

28. Davis CD. The gut microbiome and its role in obesity. Nutrition today. 2016;51(4):167.

29. Singh RK, Chang H-W, Yan D, Lee KM, Ucmak D, Wong K, et al. Influence of diet on the gut microbiome and implications for human health. Journal of Translational Medicine. 2017 Apr 8;15(1):73.

30. Vieira SM, Pagovich OE, Kriegel MA. Diet, microbiota and autoimmune diseases. Lupus. 2014;23(6):518.

31. Eshriqui I, Viljakainen HT, Ferreira SRG, Raju SC, Weiderpass E, Figueiredo RAO. Breastfeeding may have a long-term effect on oral microbiota: Results from the Fin-HIT cohort. International Breastfeeding Journal 2020 15:1. 2020 May 15;15(1):1–11.

32. Engen PA, Green SJ, Voigt RM, Forsyth CB, Keshavarzian A. the gastrointestinal microbiome: Alcohol effects on the composition of intestinal microbiota. Alcohol Research: Current Reviews. 2015 Jun 27;37(2):223.

33. Savin Z, Kivity S, Yonath H, Yehuda S. Smoking and the intestinal microbiome. Archives of Microbiology. 2018 Jul 1;200(5):677–84.

34. Erkosar B, Storelli G, Defaye A, Leulier F. Host-intestinal microbiota mutualism: "Learning on the Fly." Cell Host & Microbe. 2013 Jan 16;13(1):8–14.

35. Cabreiro F, Gems D. Worms need microbes too: Microbiota, health and aging in caenorhabditis elegans. EMBO Molecular Medicine. 2013 Sep;5(9):1300–10.

36. Wilson AS, Koller KR, Ramaboli MC, Nesengani LT, Ocvirk S, Chen C, et al. Diet and the human gut microbiome: An International Review. Digestive Diseases and Sciences 2020 65:3. 2020 Feb 14;65(3):723–40.

37. Xu Z, Knight R. Dietary effects on human gut microbiome diversity. The British journal of nutrition. 2015 Jan 1;113 Suppl (Suppl 0): S1–5.

38. Malla MA, Dubey A, Kumar A, Yadav S, Hashem A, Abd_Allah EF. Exploring the human microbiome: the potential future role of next-generation sequencing in disease diagnosis and treatment. Frontiers in Immunology. 2019;0(JAN):2868.

39. Zhu B, Wang X, Li L. Human gut microbiome: The second genome of human body. Vol. 1, Protein and Cell. Higher Education Press; 2010. p. 718–25.

40. Cani PD. Human gut microbiome: Hopes, threats and promises. Vol. 67, Gut. BMJ Publishing Group; 2018. p. 1716–25.

41. Hehemann J-H, Correc G, Barbeyron T, Helbert W, Czjzek M, Michel G. Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. Nature 2010 464:7290. 2010 Apr 8;464(7290):908–12.

42. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen Y-Y, Keilbaugh SA, et al. Linking long-term dietary patterns with gut microbial enterotypes. Science. 2011 Oct 7;334(6052):105–8.

43. Hayashi H, Sakamoto M, Benno Y. Fecal microbial diversity in a strict vegetarian as determined by molecular analysis and cultivation. Microbiology and Immunology. 2002 Dec 1;46(12):819–31.

44. Hold GL, Pryde SE, Russell VJ, Furrie E, Flint HJ. Assessment of microbial diversity in human colonic samples by 16S rDNA sequence analysis. FEMS Microbiology Ecology. 2002 Jan 1;39(1):33–9.

45. Wang X, Heazlewood SP, Krause DO, Florin THJ. Molecular characterization of the microbial species that colonize human ileal and colonic mucosa by using 16S rDNA sequence analysis. Journal of Applied Microbiology. 2003 Sep 1;95(3):508–20.

46. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalog established by metagenomic sequencing. Nature. 2010;464(7285):59.

47. Ahn J, Yang L, Paster BJ, Ganly I, Morris L, Pei Z, et al. Oral microbiome profiles: 16S rRNA pyrosequencing and microarray assay Comparison. PLoS ONE. 2011;6(7):22788.

48. Pei Z, Bini EJ, Yang L, Zhou M, Francois F, Blaser MJ. Bacterial biota in the human distal esophagus. Proceedings of the National Academy of Sciences. 2004 Mar 23;101(12):4250–5.

49. Bik EM, Eckburg PB, Gill SR, Nelson KE, Purdom EA, Francois F, et al. Molecular analysis of the bacterial microbiota in the human stomach. Proceedings of the National Academy of Sciences. 2006 Jan 17;103(3):732–7.

50. Dave M, Higgins PD, Middha S, Rioux KP. The human gut microbiome: Current knowledge, challenges, and future directions. Vol. 160, Translational Research. Mosby Inc.; 2012. p. 246–57.

51. Heinken A, Ravcheev DA, Baldini F, Heirendt L, Fleming RMT, Thiele I. Systematic assessment of secondary bile acid metabolism in gut microbes reveals distinct metabolic capabilities in inflammatory bowel disease. Microbiome 2019 7:1. 2019 Oct;7(1):1–18.

52. Lupp C, Robertson ML, Wickham ME, Sekirov I, Champion OL, Gaynor EC, et al. Host-mediated inflammation disrupts the intestinal microbiota and promotes the overgrowth of Enterobacteriaceae. Cell Host & Microbe. 2007 Oct;2(2):119–29.

53. Arrieta MC, Stiemsma LT, Dimitriu PA, Thorson L, Russell S, Yurist-Doutsch S, et al. Early infancy microbial and metabolic alterations affect risk of childhood asthma. Science Translational Medicine. 2015 Oct;7(307).

54. Ley RE. Obesity and the human microbiome. Current Opinion in Gastroenterology. 2010 Oct;26(1):5–11.

55. Ley RE, Turnbaugh PJ, Klein S, Gordon JI. Human gut microbes associated with obesity. Nature 2006 444:7122. 2006 Oct;444(7122):1022–3.

56. Sheh A, Fox JG. The role of the gastrointestinal microbiome in Helicobacter pylori pathogenesis104161/gmic26205. 2013 Aug 19;4(6).

57. Elfil M, Kamel S, Kandil M, Koo BB, Schaefer SM. Implications of the gut microbiome in Parkinson's Disease. movement disorders. 2020 Jun 1;35(6):921–33.

58. Michail S, Durbin M, Turner D, Griffiths AM, Mack DR, Hyams J, et al. Alterations in the gut microbiome of children with severe ulcerative colitis. Inflammatory bowel diseases. 2012 Oct;18(10):1799.

59. Brunkwall L, Orho-Melander M. The gut microbiome as a target for prevention and treatment of hyperglycaemia in type 2 diabetes: From current human evidence to future possibilities. Diabetologia. 2017 Jun 1;60(6):943.

60. Mu Q, Kirby J, Reilly CM, Luo XM. Leaky Gut as a danger signal for autoimmune diseases. Frontiers in Immunology. 2017 May 23;8(MAY):598.

61. Alkasir R, Li J, Li X, Jin M, Zhu B. Human gut microbiota: The links with dementia development. Protein & Cell. 2017 Feb 1;8(2):90.

62. de Palma G, Nadal I, Medina M, Donat E, Ribes-Koninckx C, Calabuig M, et al. Intestinal dysbiosis and reduced immunoglobulin-coated bacteria associated with coeliac disease in children. BMC Microbiology. 2010; 10:63.

63. Collado MC, Donat E, Ribes-Koninckx C, Calabuig M, Sanz Y. Specific duodenal and faecal bacterial groups associated with paediatric coeliac disease. Journal of clinical pathology. 2009 Mar;62(3):264–9.

64. Wacklin P, Kaukinen K, Tuovinen E, Collin P, Lindfors K, Partanen J, et al. The duodenal microbiota composition of adult celiac disease patients is associated with the clinical manifestation of the disease. Inflammatory Bowel Diseases. 2013 Apr 1;19(5):934–41.

65. Sánchez E, Donat E, Ribes-Koninckx C, Fernández-Murga ML, Sanz Y. Duodenal-mucosal bacteria associated with celiac disease in children. Applied and Environmental Microbiology. 2013;79(18):5472.

66. van Cutsem E, Sagaert X, Topal B, Haustermans K, Prenen H. Gastric cancer. Vol. 388, The Lancet. Lancet Publishing Group; 2016. p. 2654–64.

67. Catalano V, Labianca R, Beretta GD, Gatta G, de Braud F, van Cutsem E. Gastric cancer. Vol. 71, Critical Reviews in Oncology/Hematology. 2009. p. 127–64.

68. Brawner K, Morrow C, Smith P. Gastric microbiome and gastric cancer. Cancer journal (Sudbury, Mass). 2014;20(3):211–6.

69. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians. 2021 May 1;71(3):209–49.

70. Ernst PB, Gold BD. The Disease Spectrum of Helicobacter Pylori: The immunopathogenesis of gastroduodenal ulcer and gastric cancer. 101146/annurev. micro541615. 2003 Nov 28; 54:615–40.

71. Coussens LM, Werb Z. Inflammation and cancer. Nature 2002 420:6917. 2002 Dec 26;420(6917):860–7.

72. Sarhadi V, Mathew B, Kokkola A, Karla T, Tikkanen M, Rautelin H, et al. Gut microbiota of patients with different subtypes of gastric cancer and gastrointestinal stromal tumors. Gut Pathogens 2021 13:1. 2021 Feb 17;13(1):1–9.

73. Wang X, Wei M, Sun Z. An association study of histological types of gastric carcinoma with Helicobacter pylori infection. Cell Biochemistry and Biophysics 2014 70:2. 2014 Jun 5;70(2):1283–7.

74. Kunz PL, Gubens M, Fisher GA, Ford JM, Lichtensztajn DY, Clarke CA. Long-term survivors of gastric cancer: A California population-based study. https://doi.org/101200/JCO2011358028. 2012 Sep 4;30(28):3507–15.

75. Kim K-M, Kwon M-S, Hong S-J, Min K-O, Seo E-J, Lee K-Y, et al. Genetic classification of intestinal-type and diffuse-type gastric cancers based on chromosomal loss and microsatellite instability. Virchows Archiv 2003 443:4. 2003 Aug 15;443(4):491–500.

76. Vauhkonen M, Vauhkonen H, Sajantila A, Sipponen P. Differences in genomic instability between intestinal- and diffuse-type gastric cancer. Gastric Cancer 2005 8:4. 2005 Nov;8(4):238–44.

77. Oppelt PJ, Hirbe AC, Tine BA van. Gastrointestinal stromal tumors (GISTs): Point mutations matter in management, a review. Journal of Gastrointestinal Oncology. 2017 Jun 1;8(3):466.

78. Iizasa H, Ishihara S, Richardo T, Kanehiro Y, Yoshiyama H. Dysbiotic infection in the stomach. World Journal of Gastroenterology. 2015 Oct 28;21(40):11450.

79. Davies J. in a map for human life, count the microbes, too. Science. 2001 Mar 23;291(5512):2316–2316.

80. Lloyd-Price J, Abu-Ali G, Huttenhower C. The healthy human microbiome. Vol. 8, Genome Medicine. BioMed Central Ltd.; 2016.

81. Relman DA. New Technologies, Human-Microbe Interactions, and the Search for Previously Unrecognized Pathogens. The Journal of Infectious Diseases. 2002 Dec 1;186(Supplement_2): S254–8.

82. Shetty SA, Lahti L. Microbiome data science. Journal of biosciences. 2019 Oct 1;44(5):1–6.

83. Abellan-Schneyder I, Matchado MS, Reitmeier S, Sommer A, Sewald Z, Baumbach J, et al. Primer, Pipelines, Parameters: Issues in 16S rRNA Gene Sequencing. mSphere. 2021 Feb 24;6(1).

84. Degnan PH, Ochman H. Illumina-based analysis of microbial community diversity. The ISME Journal. 2012 Jan;6(1):183.

85. Lahti L, Shetty S. Introduction to the microbiome R package. http://microbiome.github.com/microbiome. 2017.

86. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, Mcglinn D, et al. Package "vegan" Title Community Ecology Package Version 2.5-7. 2020.

87. McMurdie PJ, Holmes S. Phyloseq: A bioconductor package for handling and analysis of high-throughput phylogenetic sequence data. Pacific Symposium on Biocomputing. 2012;235–46.

88. Xia Y, Sun J. Hypothesis testing and statistical analysis of microbiome. Vol. 4, Genes and Diseases. Elsevier; 2017. p. 138–48.

89. Callahan BJ, Sankaran K, Fukuyama JA, Mcmurdie PJ, Holmes SP, Lahti L, et al. Bioconductor Workflow for Microbiome Data Analysis: from raw reads to community analyses [version 2; peer review: 3 approved] report. 2016.

90. Thioulouse J. Simultaneous analysis of a sequence of paired ecological tables: A comparison of several methods. 101214/10-AOAS372. 2011 Dec 1;5(4):2300–25.

91. Youssef O, Lahti L, Kokkola A, Karla T, Tikkanen M, Ehsan H, et al. Stool Microbiota Composition Differs in Patients with Stomach, Colon, and Rectal Neoplasms. Digestive Diseases and Sciences. 2018 Nov 1;63(11):2950.

92. Sarhadi V, Lahti L, Saberi F, Youssef O, Kokkola A, Karla T, et al. Gut microbiota and host gene mutations in colorectal cancer patients and controls of Iranian and Finnish origin. Anticancer Research. 2020 Mar 1;40(3):1325–34.

93. Nguyen NP, Warnow T, Pop M, White B. A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity. Vol. 2, npj Biofilms and Microbiomes. Nature Publishing Group; 2016.

94. McMurdie PJ, Holmes S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLOS ONE. 2013 Apr 22;8(4): e61217.

95. Hughes JB, Hellmann JJ, Ricketts TH, Bohannan BJM. Counting the Uncountable: Statistical Approaches to Estimating Microbial Diversity. Vol. 67, Applied and Environmental Microbiology. 2001. p. 4399–406.

96. CHAO A, YANG MCK. Stopping rules and estimation for recapture debugging with unequal failure rates. Biometrika. 1993 Mar 1;80(1):193–201.

97. Borcard D, Gillet F, Legendre P. Community Diversity. 2018;369–412.

98. Whittaker RH. Vegetation of the Siskiyou Mountains, Oregon and California. Ecological Monographs. 1960 Jul 1;30(3):279–338.

99. Jost L. PARTITIONING DIVERSITY INTO INDEPENDENT ALPHA AND BETA COMPONENTS. Ecology. 2007 Oct 1;88(10):2427–39.

100. Tuomisto H. A diversity of beta diversities: Straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity. Ecography. 2010 Feb;33(1):2–22.

101. Lin H, Peddada S das. Analysis of microbial compositions: A review of normalization and differential abundance analysis. npj Biofilms and Microbiomes 2020 6:1. 2020 Dec 2;6(1):1–13.

102. Fernandes AD, Reid JN, Macklaim JM, Mcmurrough TA, Edgell DR, Gloor GB. Unifying the analysis of high-throughput sequencing datasets: Characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Vol. 2. 2014.

103. Qian XB, Chen T, Xu YP, Chen L, Sun FX, Lu MP, et al. A guide to human microbiome research: Study design, sample collection, and bioinformatics analysis. Vol. 133, Chinese medical journal. NLM (Medline); 2020. p. 1844–55.

104. Hood L, Rowen L. The Human Genome Project: Big science transforms biology and medicine. Genome Medicine 2013 5:9. 2013 Sep 13;5(9):1–8.

105. Fan Y, Pedersen O. Gut microbiota in human metabolic health and disease. Nature Reviews Microbiology 2020 19:1. 2020 Sep 4;19(1):55–71.

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

107. Falony G, Vieira-Silva S, Raes J. Richness and ecosystem development across faecal snapshots of the gut microbiota. Nature Microbiology 2018 3:5. 2018 Apr 24;3(5):526–8.

108. Gantuya B, el Serag HB, Matsumoto T, Ajami NJ, Uchida T, Oyuntsetseg K, et al. Gastric mucosal microbiota in a Mongolian population with gastric cancer and precursor conditions. Alimentary pharmacology & therapeutics. 2020 Apr 1;51(8):770–80.

109. Wang Z, Gao X, Zeng R, Wu Q, Sun H, Wu W, et al. Changes of the gastric mucosal microbiome associated with histological stages of gastric carcinogenesis. Frontiers in Microbiology. 2020 May 29; 11:997.

110. Wu GD, Compher C, Chen EZ, Smith SA, Shah RD, Bittinger K, et al. Comparative metabolomics in vegans and omnivores reveal constraints on diet-dependent gut microbiota metabolite production. Gut. 2016 Jan 1;65(1):63.

111. Smith CC, Snowberg LK, Gregory Caporaso J, Knight R, Bolnick DI. Dietary input of microbes and host genetic variation shape among-population differences in stickleback gut microbiota. The ISME Journal 2015 9:11. 2015 Apr 24;9(11):2515–26.

112. Fazlollahi M, Chun Y, Grishin A, Wood RA, Burks AW, Dawson P, et al. Early-life gut microbiome and egg allergy. Allergy: European Journal of Allergy and Clinical Immunology. 2018 Jul 1;73(7):1515–24.

113. Bunyavanich S, Shen N, Grishin A, Wood R, Burks W, Dawson P, et al. Early-life gut microbiome composition and milk allergy resolution. Journal of Allergy and Clinical Immunology. 2016 Oct 1;138(4):1122–30.

114. Russell JT, Roesch LFW, Ördberg M, Ilonen J, Atkinson MA, Schatz DA, et al. Genetic risk for autoimmunity is associated with distinct changes in the human gut microbiome. Nature Communications. 2019 Dec 1;10(1).

115. Stevens BR, Roesch L, Thiago P, Russell JT, Pepine CJ, Holbert RC, et al. Depression phenotype identified by using single nucleotide exact amplicon sequence variants of the human gut microbiome. Molecular Psychiatry. 2020.

116. Sze MA, Schloss PD. The Impact of DNA Polymerase and Number of Rounds of Amplification in PCR on 16S rRNA Gene Sequence Data. mSphere. 2019 Jun 26;4(3).

117. Edgar RC. Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences. PeerJ. 2018;2018(4).

118. Sierra MA, Li Q, Pushalkar S, Paul B, Sandoval TA, Kamer AR, et al. The Influences of Bioinformatics Tools and Reference Databases in Analyzing the Human Oral Microbial Community. Genes. 2020 Aug 1;11(8):1–12.

119. Quinn TP, Crowley TM, Richardson MF. Benchmarking differential expression analysis tools for RNA-Seq: Normalization-based vs. log-ratio transformation-based methods. BMC Bioinformatics. 2018 Jul 18;19(1):1–15.