METAGENOMICS ANALYSIS OF DISEASE-RELATED HUMAN GUT MICROBIOTA

A Dissertation
Presented to
The Academic Faculty

By

Congmin Xu

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
Coulter Department of Biomedical Engineering

Georgia Institute of Technology & Emory University School of Medicine

August 2020

METAGENOMICS ANALYSIS OF DISEASE-RELATED HUMAN GUT MICROBIOTA

Approved by:

Dr. Huaiqiu Zhu, Advisor
School of Biomedical Engineering
Peking University

Dr. Peng Qiu, Advisor
Department of Biomedical Engineering
Georgia Institute of Technology

Dr. Liping Duan
Department of Gastroenterology
Peking University Third Hospital

Dr. Chenggang Zhang
Institute of Biotechnology
Academy of Military Medical Sciences

Dr. Ziding Zhang
State Key Laboratory for Agrobiotechnology
China Agricultural University

Dr. Zhifei Dai
Department of Biomedical Engineering
Peking University

Dr. Jianzhon Xi
Department of Biomedical Engineering
Peking University

Date Approved: June 15, 2020

This dissertation is dedicated to the exploration of unknown unknowns.

ACKNOWLEDGEMENTS

I am deeply grateful to my advisors Huaiqiu Zhu (Peking University) and Peng Qiu (Georgia Institute of Technology). Without their guidance and persistent help this thesis would not have been possible. Dr. Zhu's coaching and support were invaluable. Dr. Qiu has been extraordinarily tolerant and supportive during my hardest time. I have benefited from both of them such a lot.

My appreciation also goes to my collaborators Xiao Guo, Xiaoqi Wang, Zhe Wang, Mo Li and Zhongjie Xie. Their assistance means a lot to this thesis and I have learned plenty of different things from each of them. A special collaborator deserves my deepest appreciation, which is my husband Dr. Kuang Chen. He supports me both technically and spiritually.

Special thanks to the financial support from China Scholarship Council and Ministry of Education of the People's Republic of China during my PhD peroid, without which I can not accomplish this thesis.

Last but not least, I owe my deepest gratitude to my mom, my dad and my elder brother. They help me build up my personality and always see the positive side even though in a bad situation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Focusing on human gut microbiota, this research work covered the most advanced aspects about the metagenomic data analysis of human gut microbiota, and explored the possibility of putting the findings about disease-related human microbiota into application.

In Chapter 2 of this dissertation, we pre-processed and carried out a uniform annotation of the raw data of human gut microbiota from hosts suffering various diseases by applying the state-of-the-art bioinformatics tools. For systematic analysis, a novel binning unit was defined, functional taxonomic unit. With the annotation result, we answered the question of how the ecological niches of gut microbiota correlate with the host health in every step of a well-designed meta-analysis, covering all the four aspects , i.e. taxonomic composition, functional carriage of these microbes, taxonomic co-occurrence network and also functional gene-gene interaction network. Universal taxonomic and functional biomarkers were identified. Interesting finding from the gene-gene interaction network and significantly alteration of taxonomic network patterns indicated that the gut microbiota inside human gut aggregate a 'super organism' and influence the host health in a community manner. In summary, taxonomic composition, microbial functions, microbial correlations and the interactions of microbial functions are four indispensable components for characterizing microbial community, which should be the comprehensive way for defining a pan-microbiome. This literal definition of pan-microbiome provides a practical framework for designing future research works.

In Chapter 3 of this dissertation, we proved the existence of an aging progression of human gut microbiota by applying unsupervised machine learning approaches on metagenomics data. We applied an unsupervised machine learning approach SPD on genera abundance profile of human gut microbiota quantified by 16S rRNA sequencing data. Without using the age information of the samples, SPD sorted sample groups on a minimal spanning tree that recapitulated the aging progression. This result indicated the existence of an

aging progression reflected in the human gut microbiota. In the meantime, we found 35 genera associated with this age-related progression. Some of these genera were not identified using the commonly-used statistical approaches for metagenomics analysis. Literature review of these 35 genera led to a lot of evidences of the functional relevance of these genera. The evidences collectively indicated an age-related decline of the beneficial functions of gut microbiota, as well as increase of inflammation and diseases, especially for the elderly people older than 90s. This is the first study characterizing the human gut microbiota in a trajectory manner, which sheds light on the possibility of exploring diverse approaches for conducting metagenomics analysis.

In Chapter 4 of this dissertation, we further explored to develop a machine-learning based tool LightCUD in Chapter 4, which was designed to assist the diagnosis of IBD based on human gut microbiome. The well-designed feature selection steps and comparison of different machine learning algorithms contributed to a high-performance tool. Regarding the high-speed development and popularity of NGS, LightCUD highlights the potential of diagnostic tools developed with machine learning algorithms based on the data of human gut microbiome.

In Chapter 5 of this dissertation, we released the first database integrating disease-related genes of human gut microbiota, named DREEM, which provides a clue and data resources for those studies about disease-related changes of gut microbiota.

CHAPTER 1
INTRODUCTION

1.1 Background

1.1.1 Human gut microbiota

The studies about the human microbiota could be traced back to 1680s, during which Antonie van Leewenhoek had compared his fecal and oral microbiota[1, 2]. He noticed the striking differences of microbes between these two habitats. Some others consider the origin of microbiota research should be two centuries later, marked by the publication of *A Flora and Fauna within Living Animals* by Joseph Leidy in 1853[3]. Then, Louis Pasteur, lie Metchnikoff, Theodor Escherich and some other pioneer microbiologists and immunologists, laid the foundations for characterizing the interactions between host and microbiota. Pasteur's medical discoveries proved the germ-theory of disease at the first time, and he also pointed out the potential value of non-pathogenic microorganisms for maintaining a normal human physiology, with a famous quote "Life would not long remain possible in the absence of microbes"[4, 5]; Metchnikoff discovered that phagocyte could kill pathogens[6]; and Escherich described *E.coli* at the first time and was convinced that the intestinal bacteria are essential to the physiology of digestion[7]. Alfred Nissle, a German physician, isolated the Escherichia coli Nissle 1917 strain from human gut in 1917 and showed the protective role of this strain against pathogens, which remains a commonly used probiotic[8]. However, 'microbiota' has not been documented as a basic microbiology term until 50 years ago, which was specified as "a catalog of microbes"[9, 10]. Latterly, the term 'microbiome' was defined in 1988 as "A convenient ecological framework in which to examine biocontrol systems is that of the microbiome. This may be defined as a characteristic microbial community occupying a reasonably well defined habitat which has

distinct physio-chemical properties. The term thus not only refers to the microorganisms involved but also encompasses their theatre of activity"[11] (which was recoined by Joshua Lederberg in 2001 as "microbiota and their genes"[12]). At the meantime, pioneer microbiologists discovered the methods to culture anaerobic organisms in the laboratory, which facilitated the understanding about the composition and function of the microbiota communities living on the surfaces of human body and how those microbiota changed throughout human lives[13]. In 1960s, the first in vivo experiment proved that germ-free mice can only recover normal physiology by being colonized with bacteria from faeces[14].

Following all these great advances in understanding microbiota, the great plate count anomaly soon became apparent in 1980s[15], that the majority of microbes from natural environment can not be cultured or observed in the lab. This observation motivated the development of sequencing-based approaches to identify unculturable microorganisms, which were first applied to study environmental microorganisms. Subsequently the sequencing-based methods were adapted to the analysis of human-associated communities, which has provided an unprecedented view into the composition and function of the human microbiota.

"When you are hungry, you are not alone. When you are sick, you are not alone, either". The human body is home for tens of trillions of microbes across different body sites, including nasal passages, oral cavity, skin, gastrointestinal tract, and urogenital tract *etc*[17](Figure 1.1). Especially, inside the human gut reside a huge amount of microbes, the number of which was estimated to be ten times the number of human body cells (Figure 1.2), and 100 times as many genes as the human genome, most of which confer physiological function. These gut microbiota are critically important to host health and have been reported as an important virtual organ[18]. In fact the structure of gut microbiota is the result of a continuous co-evolutionary history of interactions between the host bodies and intestinal microbes. This intimate association has affected both sides[19], and as a result all higher organisms negotiate a truce with their commensal microbes and battle pathogenic

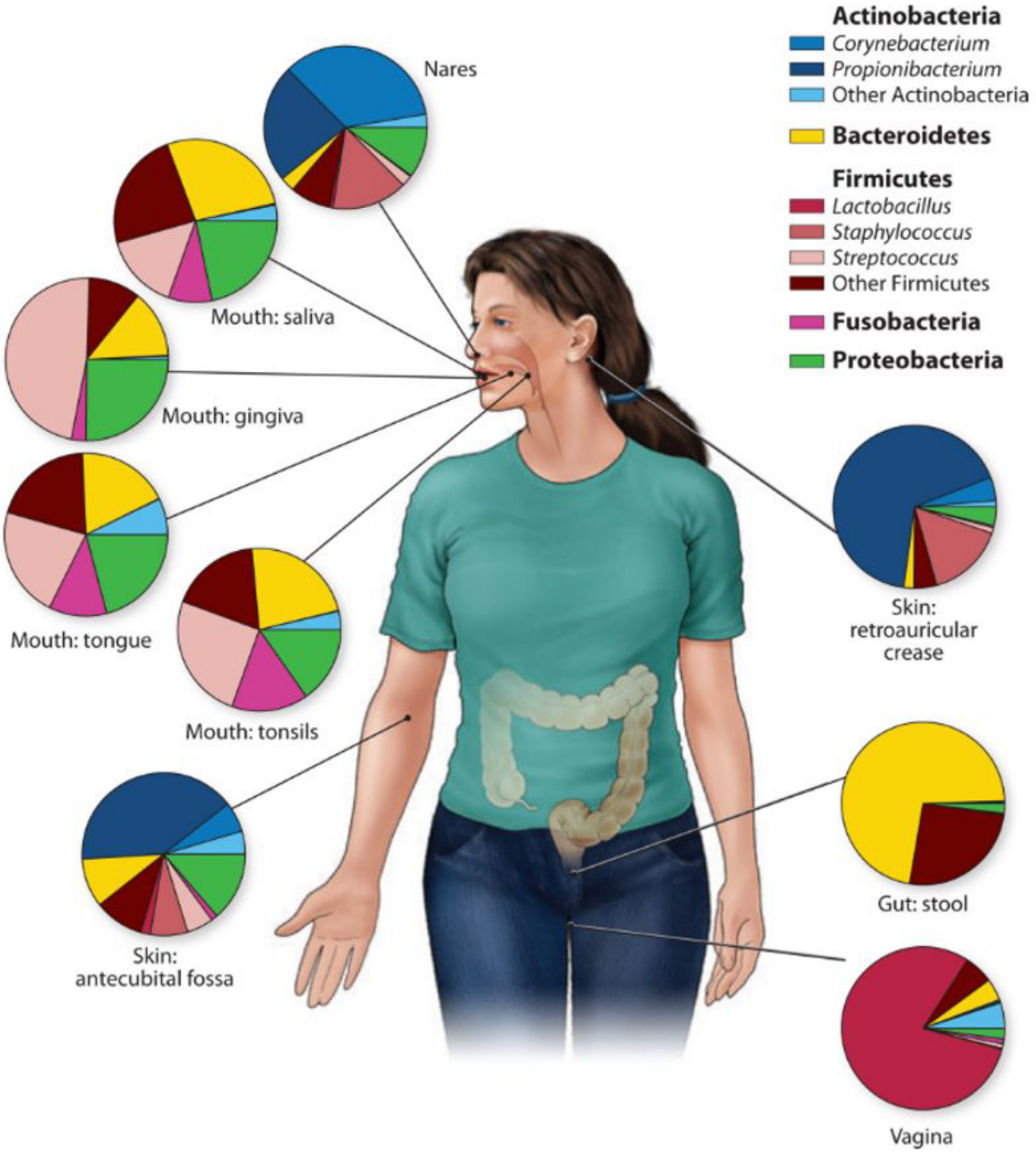

Figure 1.1: Genus- and phylum-level classification of bacteria colonizing different parts of a subject [16].

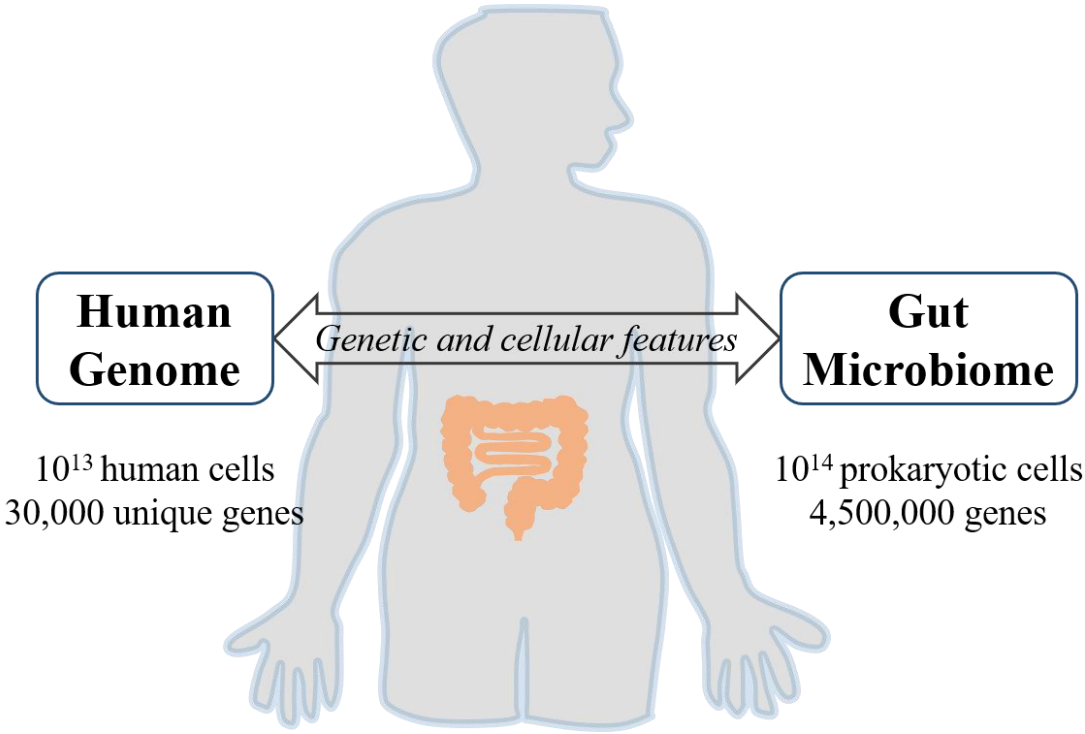

Figure 1.2: The gut microbiome is as important as human host genome.

microbes on a daily basis. Consequently, this "forgotten organ" constitute the backbone of human gut ecological system by controlling the biochemical cycling of elements essential for life. The gut microbiota benefit host through various approaches such as supplying nutrition, influencing immune system, resisting pathogens and participating in the maintenance of health[17, 20] (Figure 1.3).

1.1.2 Metagenomics analysis

In 1995, Fleischmann et al. performed whole chromosome sequencing of bacteria *Haemophilus influenzae* Rd, which released the only complete genome sequence from a free-living organism[21]. Since then, sequencing the microbial genome has become a universal means for studying microbes. As of March 2020, the National Center for Biotechnology Information has collected 245,875 complete genomes of prokaryotic microorganisms. Such massive amount of genomic data has spawned the development of comparative genomics

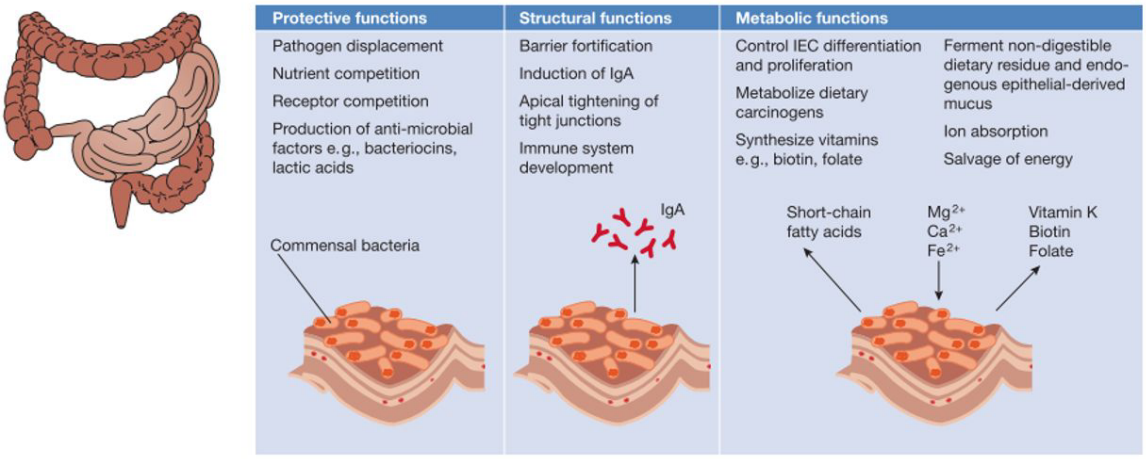

Figure 1.3: Commensal bacteria exert a miscellany of protective, structural and metabolic effects on the intestinal mucosa[18].

and systematic biology. The molecular level analysis of microorganisms have promoted our recognition about the association between genomic structure and its function. However, the culture-based single genome annotation method has great limitations. First, only less than 1% of microorganisms in the nature can be isolated and cultured[22]; Secondly, the genomes of microorganisms that are sequenced after separation and culture have a preference and cannot reflect the real situation of the original environment[23, 24]; Last but not least, microbes do not exist in isolation in a natural ecosystem (such as water, air, soil, intestines, etc.), but rather in communities in which competition and cooperation are essential for shaping the composition and function of a microbial community[25]. Since single microbial isolates in lab cultures can not accurately characterize the environmental communities, researchers have developed approaches to move beyond single pure-culture laboratory experiments to understand microbial community composition, structure, function,and evolution. The metagenomics analysis provided an effective way to overcome the above mentioned limitations[25].

Metagenomics is the study of genetic material recovered directly from environmental samples in an untargeted (shotgun) way, which facilitated the identification of microbial genome sequences from environmental samples without the need of cultivating these or-

ganisms in the laboratories[26]. As shown in Figure 1.4, the whole pipeline for metagenomics analysis includes collecting samples from the original environment, sample quality control, extracting all the DNA of all microorganisms from the samples, building up gene library and conducting whole genome sequencing/16S rRNA amplicon sequencing. Once getting the raw data, short reads were assembled into contigs, from which the relative abundance of every microbe and the potential functionality of the entire community could be referred. With metagenomics analysis, we could systematically characterize the microbial community structure, biodiversity, social relationship, co-evolution and the interaction between host and microbiota, which broad our understanding about human microobiota. The rapid development of high-throughput next generation sequencing (NGS) technology further facilitate the metagenomics analysis, but put more challenge on data analysis and integration.

Because of the experimental process, the total number of counts per sample is highly variable due to different library size. The statistical analysis of microbiome abundance data usually starts with a normalization step. The most-common used normalization method is to transfer the original raw count into ration of every count divided by the number of total reads (sequencing depth) in each sample. Some other approaches have been proposed for data normalization, for instance Aitchison's log-ratio approach, the centered log-ratio transformation (clr), and rarefaction normalization, which randomly select the same number of reads from each sample[27, 28]. Based on the normalized data, one could explore the microbiome composition to identify possible data structures. Diversity is one of the most important indicators for the quality of an ecosystem, which could be divided into two categories, alpha diversity (within sample diversity) and beta diversity (between samples diversity). Alpha diversity measures the homogeneity regarding abundance of the different species in a sample by integratin both their richness and evenness. Commonly used measures of alpha diversity include Chao1 index and Shannon index. Beta diversity characterizes the differences in microbial composition between samples, including Bray-Curtis,

UniFrac and weighted UniFrac distances. Another important step for exploring data structures is to draw ordination plots for visualizing the distances between samples. In order to draw the ordination plots, the high dimensional data need to be mapped onto tow or three dimensions while keeping the main variance as much as possible. The most commonly used methods for dimension reduction include PCA, PCoA and NMDS[29]. Subsequently, the inference analysis is performed to identify those features of interest with differential abundance regarding different host status. When one is interested in the global difference in microbial compositions, multivariate comparison like PERMANOVA[30] could be applied. Univariate differential testing could be applied when one wants to identify specific features differently distributed between samples groups. The common used methods include t-test, Wilcoxon rank-sum test and Kruskal-Wallis test. Microbes interact with each other and also with the host. The correlation analysis could be applied to identify host condition associated features or construct microbial correlation networks. The approached for computing correlation include naive correlation coefficients like Bray-Curtis distance, Pearson correlation and Spearman correlation, also some toolkits like CoNet,LSA, MIC, RMT and SparCC[31]. The metagenomics analysis include but is not limited to aforementioned approaches. Actually, it's far from enough and there is still considerable need for improvement in current technologies and exploration of new methods.

1.1.3 Disease-related variability of human gut microbiota

Regarding this rapid progress, the metagenomics analysis of the NGS data has been playing an important role in understanding the impact of gut microbial ecosystem on human diseases. In the past decade, the correlation between gut flora and the health and disease of host has become the hotspot of researches, remarkably further driven by the launch of the National Microbiome Initiative in 2016, Human Microbiome Project (HMP) and iHMP project targeting multiple omics technologies. Through case-control metagenome-wide association studies, the population structures of gut flora have been well studied. The clinical

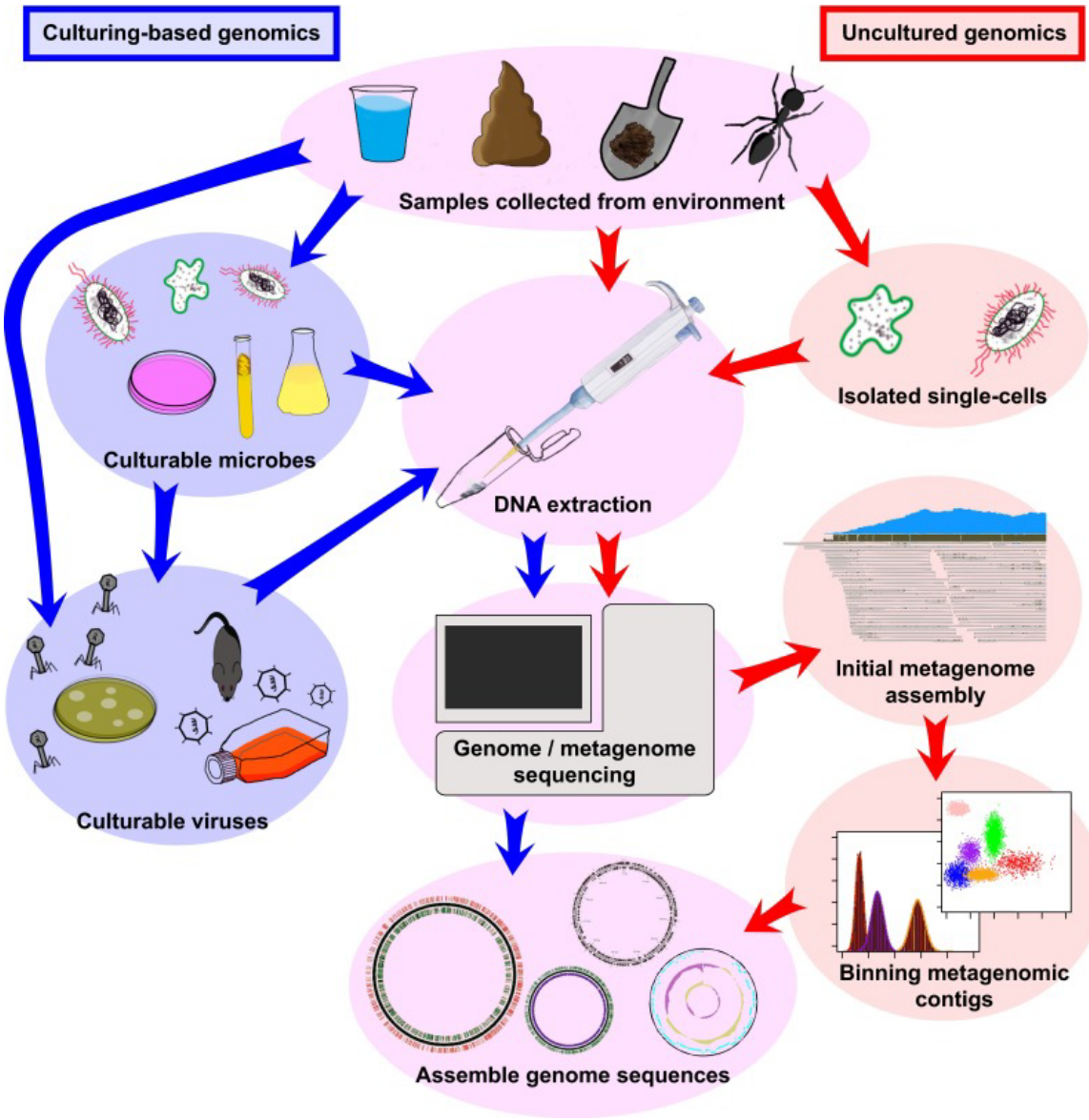

Figure 1.4: Illustration of simplified pipelines to obtain genome sequences from cultured and uncultured microbes[26].

trials and animal experiments have revealed that the alteration of the gut microbiota are closely associated with the happen of diseases like type II diabetes (T2D)[32, 33], Crohn's diseases[34], obesity[35], depression[36] and colorectal cancer[37], et al., illustrated by obvious changes in community structure and metabolic potential.

Using 16S rRNA sequencing data, IBD-affected individuals were reported to have 30-50% reduced biodiversity of gut microbes[38]. Studies focusing on bacterial 16S rRNA genetic phylotypes suggested significant phylotypic alterations in the intestinal microbiota of irritable bowel syndrome (IBS) patients[36, 39]. It has been proved that the relative portion of *Bacteroidetes* decreased in obese people, while increased with weight loss on two types of low-calorie diet[40]. Gut microbiota from inflammatory bowel disease (IBD) patients were detected to produce significantly more short-chain fatty acids and ammonia than that from healthy individuals[41]. Also, studies of depression have demonstrated the overrepresentation of *Bacteroidales* and underrepresentation of *Lachnospiraceae*[42]. These studies indicated that the gut flora are strongly associated with human health taxonomically, therefore lead to a stirring of interests of variation in the gut microbiota.

With whole genome sequencing data, the disease-related alteration of human gut microbiota could be characterized not only on species level, but also on gene level. An obesity-related study demonstrated that the obesity-associated signals originated from the host gut microbiome may be much stronger than that from the host genome[35]. A T2D-related study showed that the T2D-enriched microbial markers involve membrane transport of sugars, branched-chain amino acid transport, methane metabolism, xenobiotics degradation and metabolism, and sulphate reduction[33]. It is also documented that from the point of gut flora, liver-cirrhosis was associated with assimilation or dissimilation of nitrate to or from ammonia, GABA, phosphor transferase systems, haem biosynthesis and some types of membrane transport[43]. 15,894 genes were indicated as the significantly different functional genes, and they could also be applied to an efficient discrimination of lean and obese individuals[35]. Another study indentified 75,245 genes alternated between patients

with liver cirrhosis and health controls[43]. These WGS studies of gut microbiome have exhibited a lot of disease-related genes and metabolic pathways of gut flora.

Intra-individual changes have also been revealed as driven by restoring forces within a dynamic system though the composition of the adult gut microbiota appears to be relatively stable compared to the inter-individual changes[44]. Aging is a process capturing many aspects of the biological variation of the human body, which was accompanied by an increased incidence of infection and functional decline in the gut of elderly individuals[45]. Several previous studies have reported age-related changes of human gut microbiota[46, 47, 48, 49, 50, 51, 52, 53, 54]. By culturing microbes, Hopkins et al. found larger number of *Enterobacteria* in children's fecal than in adults[46]. Yatsunenko et al. found the number of *Bifidobacterium* declined as ages of the hosts increased using 16S rRNA sequencing[48]. Odamaki et al. revealed that there was an increasing proportion of *Bacteroides*, *Eubacterium* and *Clostridiaceae* accompanying aging; while *Enterobacteriaceae* were enriched in elderly and infant; *Bifidobacterium* were more abundant in infants; *Lachnospiraceae* were more abundant in adults[47]. Stewart et al. discovered L-lactate dehydrogenase major in milk fermentation declined and transketolase major in the metabolism of fiber increased over the first years of life[50] using whole genome sequencing.

1.1.4 Clinical applications

Regarding all these advances, it comes to a common sense that maintaining gut microbial structure and function is beneficial to human health. Normal gut bacteria as a union, their metabolites are essential for host physiologic activities. The dysbiosis of gut microbiota has been associated with a lot of infectious, inflammatory, functional, and nutritional pathological conditions. Manipulating the gut microbiota in a well-designed approach could help prevent or treat some diseases.

- Pathogen target treatment

Antibiotics has been used to kill specific pathogen and modulate the dysbiosis of human gut microbiota since 1920s, although they can disrupt the stability of microbial community at the same time. But the beneficial clinical outcome is obvious and has made antibiotics an established and effective treatment for a number of infectious intestinal diseases, like infectious *C. difficile*-associated diarrhea. Even for those diseases (e.g., IBD and IBS) without a clear recognition of pathogen-associated etiology, antibiotics treatment are also one of the most common used treatment approaches. Mouse model studies have also proved the potential reversing or attenuating effect of gut microbiota on the dysbiosis of gut microbiota[55]. However, the concerns should be addressed whenever using antibiotics for treating diseases, since it is in the risk of resulting in an increased susceptibility to diseases and the dysregulation of host immune homeostasis.

- Probiotics modulation

Probiotics are defined as a set of live microorganisms conferring a health benefit to the host by improving or restoring the gut flora. Probiotic therapy could be dated back to almost 100 years age, when Elie Metchnikoff (this name appears twice in this chapter) suggested that the yogurt consumption make Bulgarian peasants live longer. In both China and Japan, fermentation using various microorganisms is a traditional method to produce miso (soybean paste), sake (wine made from rice), natto (fermented soybean), pickles, fermented dairy products, and many other products[56]. Probiotics have been studied in recent years as an approach for modulating microbial populations and functions in order to promote host health and manage or prevent intestinal diseases. Ingestion of probiotics has been used to treat a lot of pathological disorders, for instance allergic reactions, constipation, infections in infancy, and IBS. A balanced enteric flora might competitively exclude possible pathogens, promote the intestinal immune system, and produce beneficial metabolites such as short-chain fatty acids, vitamins, amino acids like arginine, cysteine and glutamine, polyamines, antioxidants, and growth factors[57].

- Fecal Transplantation

Fecal microbiota transplantation (FMT) is a procedure to transfer slightly processed feces from a healthy donor to a recipient with some kind of conditions, with the aim to establish a healthy diverse microbial community within the gut of patients. FMT has a long history, which could date back the fourth century. In ancient Chinese medicine Ge Hong used 'yellow soup' to treat food poisoning and severe diarrhoea. In the sixteenth century, Li Shizhen used a 'soap' mixed with fresh, dry or fermented stool as oral administration to treat abdominal diseases. In seventeenth century, Fabrizio from Italy and Paullini from German reported the use of FMT. The first record about the use of FMT in Western medicine was published in 1958. Ben Eiseman and his colleagues treated patients suffering from pseudomembranous colitis, before *Clostridioides difficile* was discovered as the real cause. Nowadays, FMT is still a highly effective way for treating recurrent *Clostridium difficile* infections. With the rapid updates about the role of gut microbiota in modulating host health, FMT has been experimentally validated as efficient at treating various conditions, such as Crohns disease, ulcerative colitis, metabolic and autoimmune diseases, autism, Parkinsons disease, multiple sclerosis, irritable bowel syndrome, and chronic fatigue syndrome. The implementation of FMT is easy, and it has been proved as a cheap and reliable treatment approach[58]. However, there are still concerns about its long-term risks and the standard application protocols have not yet been established.

- Machine learning based algorithms for disease diagnosis

Machine learning approaches have been applied to conduct diseases prediction with various biological data[59]. The gut microbiome is one of the representative data type. Recently, studies have begun to predict host conditions by exploring the power of machine learning algorithms applied on human gut microbiome patterns[60]. The main challenge for implicating this kind of algorithms is that the microbial imbalance associated with disease-related dysbiosis could be caused by a wide range of reasons, which putting a

lot of challenge on feature selection. Additionally, individual studies usually have limitations like small sample sizes, various processing procedure, and inconsistent findings[61]. This situation led to the difficulty for generalizing prediction models across studies[62]. A few Meta-analysis has been performed to address these issue, combining sample cohorts from multiple microbiome studies. Duvallet et al. performed a meta-analysis covering 10 diseases to find consistent patterns characterizing disease-associated microbiome changes[61]. Pasolli et al. also performed a meta-analysis collecting 2,424 publicly available samples[62]. The authors pointed out that the addition of healthy controls from other studies to training sets could improve the disease prediction capabilities of models. The data integration and external validation could improve the robustness and generalization of models for prediction, compared to those models that are validated internally only, which call for an integrated database including disease-related markers of human gut microbiota as inclusive as possible.

1.2 Outline of this dissertation

The human gut microbiota has been revealed as an essential partner for human health. Numerous studies have been conducted around human gut microbiota by applying metagenomics analysis on data generated by next-generation sequencing data which has enhanced the understanding about the microbial compositions and functional carriage of human gut microbiota. The aims of these studies included but not limited to those studies tried to reveal the microbial diversity, functional dysbiosis, mcriobial pathways, novel genes, antibiotic resistance gene, co-evolution of host and microbiota, interaction between host and microbiota (Figure 1.5). These studies led to the revolution about the recognition of disease-related changes of human gut microbiota, which has been one of the hotspots. Those related studies have nonetheless provided important insight into the association between host diseases and gut microbial dysbiosis, there is an urgent call to integrate those scattered datasets about multiple diseases, which will assist the community to see a forest instead of

individual trees for characterizing the gut microbial community. Besides data integration, new machine learning methods and different observation angles are also in an urgent call and deserve exploration. With all these established discoveries and publicly available WGS data of human gut microbiota, developing new tools and building up databases to put those disease-related markers into application are also necessary. These are the essential parts for characterizing human gut microbiota using metagenomics analysis.

As shown in Figure 1.6, this dissertation characterized the human gut microbiota from four aspects, mainly focusing on the metagenomics analysis of disease-related human gut microbiota. The first part (referring to Chapter 2) characterized the consistent changes of human gut microbiota behind multiple diseases using case-control comparative analysis. The second part (referring to Chapter 3) characterized the dynamic changes of human gut microbiota during host aging progression in a trajectory way. The third part (referring to Chapter 4) explored the application of disease-related human gut microbiota, building up a disease diagnosis tool through applying machine learning algorithm on human gut microbiome. The last part (referring to Chapter 5) constructed a database integrating disease-related marker genes in human gut microbiome, which will benefit the community who are also interested in doing analysis or developing application tools of human gut microbiome.

- Consistent changes of human gut microbiota behind multiple diseases.

The first part (Chapter 2) of this dissertation characterized the alteration of human gut microbiota behind multiple diseases through a pan-microbiome analysis.

Variation of the human gut microbiome in different population related to various diseases has been researched by a plethora of studies[63, 17, 64, 65]. Meanwhile, increasing amount of whole-genome metagenomic sequencing of the human gut microbiome have provided important insights[66, 17, 67, 68, 69]. However, there is still a lack of comprehensive understanding on how the transformation of host pathological conditions are associated with the dysbiosis of the human gut microbial community, thus encouraging us to carry out the comprehensive analyses on the WGS data of human gut microbiome with

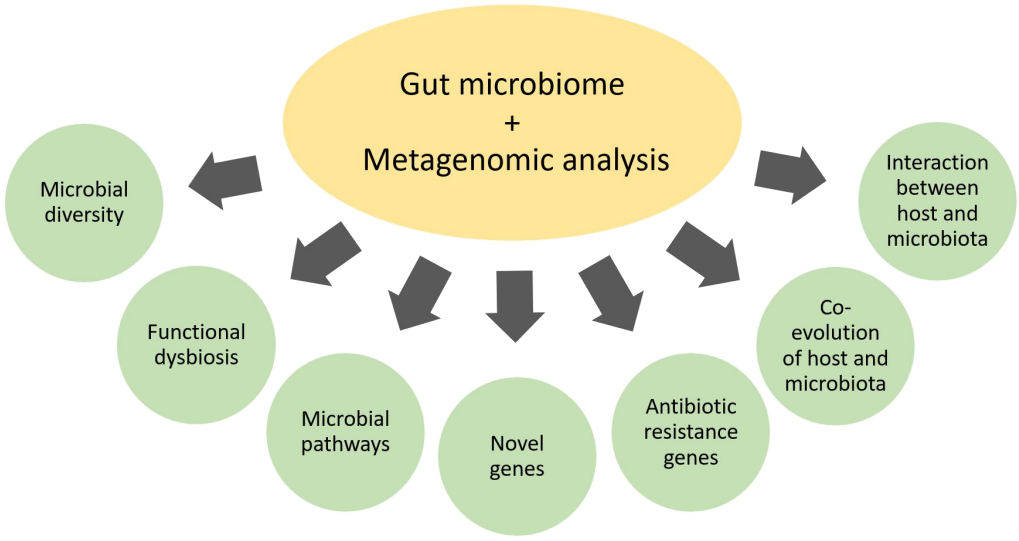

Figure 1.5: Application of metagenomics in the human gut microbiome.

regard to the hosts suffering from different kinds of diseases and dip into the question about what kinds of human gut microbial community are healthy.

One straightforward approach to answer this question is to identify microbes consistently altered in different diseases. Claire et al. tried to find the genera related to multiple diseases using 16S RNA sequencing data[70, 71, 72]which has nonetheless provided important insight. Since microbes share metabolites with the host and also among themselves[73], functional characterization of the human gut microbiome is another foremost aspect. The increasing amount of WGS data of the human gut microbiome[66, 17, 74, 75, 76] gave us a chance to make it happen. However, it could be very difficult to absolutely say some kinds of microbes or functional genes are harmful while others are beneficial, because a very large proportion of harmful microbes or functional genes in one disease might be beneficial in another disease[71, 72, 77]. The microbial community inside our gut is an ecological community of complex microbial niches and also can be considered as a super-organism. Inspired by the significance of species-species collaboration network in an ecological system and gene-gene interaction network inside the genome of an organis-

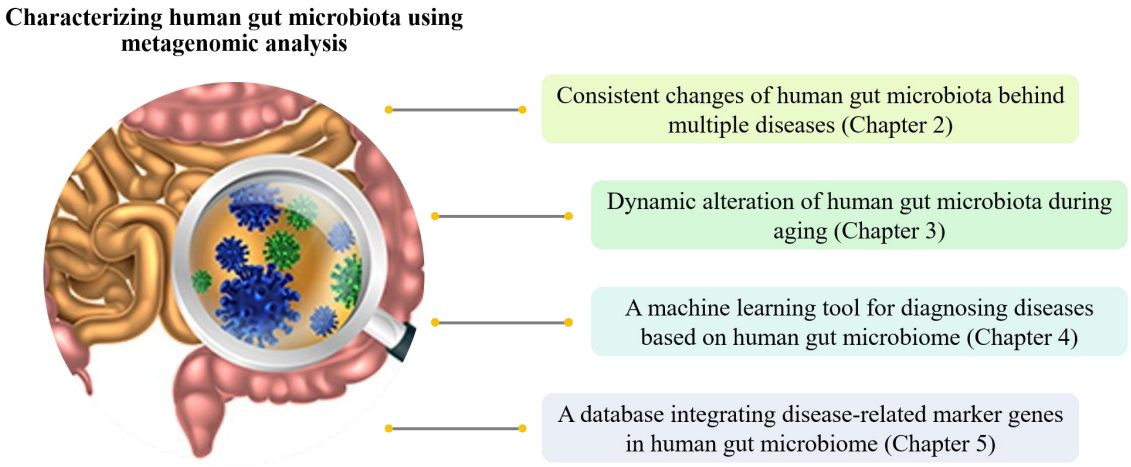

Figure 1.6: Outline of this dissertation.

m, models to characterize the microbial cooperation and also the functional genes' linkage will be super interesting and rise the chance for more systematic feature detection.

Herein, we conducted pan-microbiome analysis on a large scale of gut metagenomic WGS data collected from groups of hosts under six types of pathological conditions, T2D, Crohn's diseases (CD), ulcerative colitis (UC), liver cirrhosis, obesity and symptomatic atherosclerosis, as well as their healthy controls[33, 32, 43, 35, 78, 79, 80]. We pre-processed and carried out a uniform annotation of the raw data. With the annotation result, we carried out a comprehensive pan-microbiome analysis of human gut microbiota covering all the four aspects mentioned above, i.e. taxonomic composition, functional carriage of these microbes, taxonomic co-occurrence network and also functional gene-gene interaction network. We answered the question of how the ecological niches of gut modulate human health in every step of this well-designed meta-analysis.

- Dynamic Alteration of disease-related microbiota during aging progression.

Moving beyond Chapter 2, Chapter 3 explored to characterize the human gut microbiota in a different perspective, capturing the transformation of human gut microbiome in

a trajectory way during host aging, and characterizing how the disease-related microbiota altered during this aging progression.

Age-related changes of human gut microbiota have been revealed by several previous studies[49, 52, 51, 45, 46, 53, 50, 54, 48]. Hopkins et al. found higher numbers of *Enterobacteria* in children's fecal than adults through culturing microbes[46]. Using 16S rRNA sequencing, Yatsunenko et al. found *Bifidobacterium* declined with increasing ages[48]. Odamaki et al. revealed that aging was accompanied by increasing proportion of *Bacteroides*, *Eubacterium* and *Clostridiaceae*; *Enterobacteriaceae* were enriched in infant and elderly; *Bifidobacterium* were enriched in infants; *Lachnospiraceae* were enriched in adults[47]. Using whole genome sequencing, Stewart et al. discovered decline of L-lactate dehydrogenase (milk fermentation) and increase of transketolase (metabolism of fiber) over the first year of life[50]. In these studies, various supervised machine learning methods have been applied, including multi-group comparative analysis with permutational analysis of variance (PERMANOVA)[46, 47, 49, 50], Spearman rank correlation and Random Forest[48], as well as frequency-inverse document frequency and minimum-redundancy maximum-relevance[51], which effectively identified taxonomic or functional signatures showed aging-related changes of gut microbiota.

In this part, we proposed to explore an unsupervised machine learning approach for identifying aging-related progression of microbiota community and bacteria genera associated with this progression. The unsupervised algorithm adopted here is called Sample Progression Discovery (SPD), which was developed to identify progressive changing patterns of gene expression that reflect the biological progression in various biological processes and systems[81]. This idea was first applied to microarray gene expression analysis[81], and then extended to flow cytometry[82] and single-cell RNA-seq analysis[83]. Here, we applied SPD on community profiles extracted from 16S rRNA sequencing data of human gut microbiota samples in various age periods ranging from new-born babies to centenarians. SPD recapitulated the underlying aging progression of the data in an unsupervised

fashion, and sorted the gut microbiota samples in an order consistent to the host ages. In addition, SPD identified bacteria genera associated to the aging-related progression of gut microbiota. These findings demonstrated the existence of an aging progression of human gut microbial community, and points to important bacteria genera that characterize the aging of gut microbiota.

- A machine learning tool for diagnosing diseases based on human gut microbiota.

Moving from analysis and characterization of disease-related human gut microbiota, in Chapter 4 we explored to develop a disease discrimination tools based on the gut microbiome using disease-related marker species or genera as features for building machine learning models.

IBD is a group of inflammatory conditions of the colon and small intestine that affects over 2.5 million Europeans[84] and 3.1 million Americans[85], and has a notably increasing prevalence in the Asia-Pacific region[86]. An early accurate diagnosis can help clinicians to improve treatment. However, there is no gold standard diagnosis for monitoring quiescent disease in patients with IBD. Moreover, the two major types of IBD, UC and CD[87], have different mechanisms of tissue damage[88], necessitating different treatment strategies. It is clinically critical but usually difficult to identify the specific types of IBD, because there are no golden biomarkers or clinical tests capable of discriminating CD from UC patients in practice[89]. Even colonoscopy may miss inflammation in some parts of the gastrointestinal tract[90].

Keeping this question in mind, we developed the tool LightCUD for discriminating UC and CD from non-IBD colitis using the human gut microbiome. LightCUD embodies four high-performance modules, namely, WGS-based health vs IBD module, WGS-based UC vs CD module, 16S-based health vs IBD module and 16S-based UC vs CD module. Each module is composed of a machine learning model and a customized reference database. In details, we used the high-throughput WGS data to analyze the microbial composition of gut

microbiota samples. These samples were from patients with UC and CD, and healthy controls. The taxonomic profiles of these samples were obtained as feature abundance matrices at strain level for two WGS-based modules and at genus level for two 16S-based modules respectively. We designed a feature selection strategy for all the modules. Also, we compared the performances of five different machine learning algorithms, i.e., logistic regression, random forest, gradient boosting classifier, support vector machine and LightGBM for training each model of corresponding module[91, 92, 93, 94]. The LightGBM-based models performed best. As a result, we established four high-performance lightGBM-based modules, namely, WGS-based health vs IBD module, WGS-based UC vs CD module, 16S-based health vs IBS module and 16S-based UC vs CD module. For the two WGS-based modules, we further optimized the feature/strain sets to improve the modules performance. The result illustrated that 49 strains for WGS-based health vs IBD module and 12 strains for WGS-based UC vs CD module could achieve the best performances. Finally, we constructed and released the tool LightCUD. With 16S rRNA sequencing or WGS data from individual gut microbiota samples as input data, LightCUD predicts the probability of having IBD, and the sample identified as IBD will then be classified as UC or CD.

- A database integrating disease-related marker genes in human gut microbiome.

To benefit the community who are interested in characterization or developing application tools of human gut microbiota, we released a database of Disease-RElatEd Marker genes (DREEM database) in human gut microbiota as Chapter 5 of this dissertation.

Although some a set of microbial genes may be recognized as those associated with a specific disease, we probably just haven't seen the forest for the trees. There still lacks a common set of microbial genes related to human health and diseases in general terms. People constructed such as the IGC database, an integrated gene catalogue of intestinal microbiome, however it is not associated with specific pathogenicity and host health[66]. Therefore, an integrated general knowledge of these microbial genes certainly facilitate understanding the correlation between gut flora and human health and disease, as well as

the mechanisms of how gut flora contributes to disease process. Fortunately, large amounts of metagenomic data by current studies have been released at the public databases GenBank[95] and EMBL[96], though providing a provisional knowledge with fragmentary evidences, leading to the chance for information integration.

Herein we constructed a comprehensive database, named DREEM, which have retrieved a large scale of WGS data of human gut metagenomes, covering six types of pathological conditions, i.e., T2D, CD, UC, liver cirrhosis, symptomatic atherosclerosis and obesity. The short reads with the size of 18.63T consisting of 1,729 samples were processed with a standard procedure, involving the state-of-the-art bioinformatics tools and statistical analysis. Then we picked out 1,953,046 non-redundant DREEM genes. The DREEM genes specific to a certain disease were also stored as an individual gene set with respect to six diseases considered in the current program. Furthermore, we provided a set of Core-DREEM genes, which are shared among the samples of five metabolic syndrome related co-morbidities: T2D, CD, UC, liver cirrhosis and obesity. All DREEM genes were analyzed for taxonomic classification and functional annotation. Serving as the integration of gut microbial pathogenic gene catalogues, as a result, DREEM could be employed to detect functional and metabolic disturbance of host gut microbiomes, thus may provide brand-new strategies for host disease diagnosis and facilitate studies on human gut flora.

CHAPTER 2
CONSISTENT CHANGES OF HUMAN GUT MICROBIOTA BEHIND MULTIPLE DISEASES

2.1 Introduction

The gut microbiota has been reported to be serving as an important "forgotten organ" which embodies 100 trillion microbes[97]. The numerous microbes constitute the backbone of human gut ecological system by controlling the biochemical cycling of elements essential for life. This large and dynamic bacterial community can benefit the host through various approaches such as supplying nutrition, influencing immune system and resisting pathogens[63]. Variation of the human gut microbiome in different population related to diseases, ethnic factors, diet habitat and age has been researched by a plethora of studies[63, 17, 64, 65]. Meanwhile, increasing amount of whole-genome metagenomic sequencing of the human gut microbiome have nonetheless provided important insights[66, 17, 67, 68, 69]. However, there is still a lack of comprehensive understanding on how the transformation of host pathological conditions are associated with the dysbiosis of the human gut microbial community.Our understanding of the underlying mechanisms is still limited by its individuality and complexity.

Fortunately, it has received increasing concerns on the gut microbiota of hosts suffering from various diseases such as type 2 diabetes (T2D), Crohn's diseases, obesity, liver cirrhosis, irritable bowel syndrome (IBS) and depression[33, 32, 35, 79, 78, 43, 36]. Together, these studies revealed the compositional changes in gut microbial community of individuals with different diseases, suggesting the correlation between states of the individual health and the microbial dysbiosis.Significant discrepancies in the structure of the human gut microbiome and metabolic potential between cases and healthy controls have been re-

vealed.Despite being crucial for understanding the relationship between the gut microbial ecosystem and the host pathological conditions, these separate studies and incomparable data sets missed the forest for the trees, thus encouraging us to carry out the comprehensive analyses on the whole genome sequencing (WGS) data of human gut microbiome with regard to the hosts suffering from different kinds of diseases.

To achieve this goal,we carried out a uniform processing and analysis on a large scale of gut metagenomic WGS data collected from groups of hosts under six types of pathological conditions, T2D, Crohn's diseases, ulcerative colitis, liver cirrhosis, obesity and symptomatic atherosclerosis, as well as their healthy controls[33, 32, 35, 79, 78, 43, 80]. From this we obtained a large and comprehensive pan-microbiome of human gut microbial community regarding both bacterial members and gene functions. We then identified a set of core strains that can be used as signature taxa and elucidated the adaptive mechanism at community level by discriminating the distinct patterns of species-species correlation network between cases and controls.

2.2 Data and methods

2.2.1 Dataset collection

We collected WGS data of human gut microbiome from seven separate studies[33, 80, 32, 35, 79, 78, 43], which consists of samples with six types of diseases (T2D, obesity, Crohn's disease, ulcerative colitis, liver cirrhosis and atherosclerosis) and their corresponding control. The hosts of the samples are of widespread geographical originations (*i.e.* Europe, America and China). Detailed hosts originations and sample sizes with corresponding research publications were summarized in Table 5.1 and Table 5.2. Samples were carefully selected according to health conditions of their host recorded in original literatures. Those with comorbidity were excluded from further analysis. In general, 11.5 billion paired-end short reads of 1,729 samples, with a total size of 18.63 tera base (Tb), were downloaded from GenBank[98] and EMBL[99]. All samples were sequenced in Illumina with read

length ranging from 75bp to 102bp as recorded[33, 80, 32, 35, 79, 78, 43]. For samples from Nielsen and Colleauge's obesity study[78], we selected the healthy individuals with $BMI < 25$ as controls (63 in total) for obesity, Crohn's diseases and ulcerative colitis. The obese only individuals ($BMI \geq 30$) in healthy healthy conditions were chosen as obesity samples, of which the ones with inflammatory bowel disease were excluded. Finally, a total of 1,659 samples were included in our study. All WGS data was retrieved and processed in a standard workflow as shown in Figure 2.1a and described in details as below.

2.2.2 Short reads assembly and gene annotation

Raw short reads were assembled into contigs by InteMAP, an integrated metagenomic assembly pipeline designed for NGS sequencing data[100]. 88.75% of the short reads were assembled with the average contig length of 1,531 bp. Short reads with low quality and contigs with low sequencing depth were strained off as quality control for further analysis. Genes identification was carried out on contigs by two metagenomic gene predictors, MetaGeneMark[101] and MetaGUN/MetaTISA[102, 103], with combining genes detected by both tools to include more protein coding genes. Herein both InteMAP and MetaGUN/MetaTISA were developed by the authors[100, 102]. In total, we identified 639.3 million genes for all samples. A strategy similar to MetaHIT was implemented to construct the non-redundant gene set (the 95% identity and 90% coverage criteria), with the pairwise alignment tool replace by CD-hit[43, 104]. The obtained non-redundant gene set consists of $35,714,294$ representative genes.

2.2.3 Taxonomic classification and functional annotation

In order to obtain more accurate taxonomic classification and avoid deviation caused by the similarity between homologous genes, only contigs longer than $1,000$ bp were considered for phylogenetic assignment. A combination of a composition-based method and a sequence alignment algorithm mapping against $2,712$ genomes from NCBI RefSeq were

employed to conduct the classification[105, 106], during which alignment result was decided by the one with higher alignment score. Taxonomic information was then assigned to the representative genes of the non-redundant gene set, resulting in $2,661$ strains, $1,478$ species, 696 genera and 39 phylum , and covering $60.14\% \pm 11.68\%$ of the sequencing reads in all samples. We ensured that comparative analysis using these procedures was not biased by data set origins. Function annotation was further performed for all representative genes via BLAST[106] against COG (Cluster of Orthologous Groups of proteins) and KEGG (Kyoto Encyclopedia of Genes and Genomes) databases with the criteron of e-value$\leq 10^{-3}$[107, 108]. 42.27% genes of the non-redundant gene set were classified into $4,786$ COGs[107, 106], and 21.06% genes of the non-redundant gene set were classified into 145 KEGG pathways. Genes in the same orthologous group of COG are functionally conserved to each other when translated into proteins. In this study, we identify FTUs at a resolution considering functional orthologous as genes classified into the same COG[106, 107] and taxonomical unit of genes aligned onto the same reference genomes of one strain in NCBI[105, 109].

2.2.4 Construction of FTU abundance matrix and statistical analysis

Since bacteria can exchange genetic material, we proposed a generalized concept, namely functional taxonomic unit (FTU). An FTU is defined as a group of genes in the same cluster of functional orthologous with identical taxonomic assignment under a certain rank (i.e., phylum, class, order, family, genus, species and strain). As a basic unit of metagenomic analysis, FTU was introduced here to characterize not only the profile of signature microbes and the functional carriage of metabolic pathways, but also the linkage of these two metagenomic determinant factors. FTU is somehow analogous to OTU, which is commonly used as pragmatic proxies for microbial "species" at different taxonomic levels during 16S rRNA analysis. Using FTU we can structurally organize the abundant metagenomic data and reduce the magnitude of complexity, further conduct analysis in a systematical

manner.

Implemented via the alignment tool bowtie2[110], the occurrence of each FTU was counted by the number of mapped reads to the representative genes, followed by a normalization over gene length to eliminate bias. The occurrences of all FTUs then formed into an abundance matrix for further analysis. To exclude random mapping, only FTUs with relative abundance above 10^{-8} were considered as presented in a sample, where the relative abundance of a FTU was defined as its abundance normalized over the abundance summation of all FTUs in this sample.

In this study, genes in the same FTU were functionally conserved (classified to the same COG) and from the same genome of one strain. The relative abundance matrix and relative frequency matrix of phylums, genera, species, COGs and COG categories were constructed by the same approach as of FTUs. Wilcoxon rank-sum test was used to identify weather the features (i.e. phylums, genera, species, COG category and COG class) were significantly differing between cases and their controls with Q value thresholds (adjusted from P value to control the false discovery rate). As a nonparametric test approach, it was widely used to find out features with significantly different distribution between two sample groups using magnitude-based ranks[111].

In order to figure out how the forementioned features correlated with each other, we computed Spearman correlation coefficients with SparCC[112]. The credible correlations ($P < 0.05$) were picked to calculate the network complexity, node active index in the network. The visualization of the interaction network was created by CytoScape[113].

2.2.5 Determination of core strains and discrimination model based on lightGBM

As shown in Figure 2.2a, all six sample groups contributed to a set of core FTUs, $1,095$ FTUs for cases and $12,494$ for controls. By tracing back the taxonomic information of core FTUs, we obtained a list of shared strains for cases and controls separately. All 40 shared strains of cases were also included in the shared strain list of controls. In respec-

t to the strain relative abundance, they all ranked in top 100 abundant taxa and made up 41.72% of the community. With regard to the fact that about 40% of our data were taxonomically unclassified, the 40 core strains (40 out of $2,661$ strains) dominated 69.37% of the community Table 2.1. With the relative abundance of the 40 common features across all samples, we trained a discrimination model using lightGBM algorithm, a fast, distributed and high performance gradient boosting framework based on decision tree[91]. We evaluated its performance by a tenfold cross-validation approach and scored the discrimination ability in a receiver operating characteristic (ROC) analysis. The discriminative ability of the model was computed as the area under the ROC curve. As shown in Figure 2.2b, cases were correctly identified from controls with an average AUC of 78.68%, suggesting the 40 strains can be used as powerful biomarkers.

Table 2.1: The 40 shared strains

40 shared strains	Abundance Order	Feature IS
Mycobacterium_smegmatis_str._MC2_155	3	181
Buchnera_aphidicola_str._Ak_(Acyrthosiphon_kondoi)	45	159.8
Methylobacterium_nodulans_ORS_2060	13	152.1
Pseudomonas_putida_NBRC_14164	6	135.1
Legionella_pneumophila_2300/99_Alcoy	40	134
Streptococcus_pneumoniae_SPN034156	19	128.3
Coprococcus_catus_GD/7	2	128
Eggerthella_sp._YY7918	9	126
Nitrosomonas_sp._AL212	20	124.4
Burkholderia_gladioli_BSR3	15	123.9
Bacillus_thuringiensis_YBT-1518	7	121.9
Terriglobus_roseus_DSM_18391	53	121.1
Carnobacterium_sp._17-4	51	120

Table 2.1 continued

Desulfobulbus_propionicus_DSM_2032	8	116.3
Lactococcus_lactis_subsp._cremoris_A76	14	115.4
Desulfosporosinus_orientis_DSM_765	17	115.4
Alteromonas_macleodii_str._'English_Channel_615'	12	114.2
Escherichia_coli_O83::H1_str._NRG_857C	5	112.7
Staphylococcus_saprophyticus_subsp. saprophyticus_ATCC_15305	36	112.5
Streptococcus_pneumoniae_gamPNI0373	48	109.9
Bartonella_quintana_RM-11	28	109.3
Cyanothece_sp._PCC_7822	4	108.8
Candidatus_Nitrospira_defluvii	29	107.3
Marinobacter_hydrocarbonoclasticus_ATCC_49840	22	105.7
Blattabacterium_sp._(Blaberus_giganteus)	85	105.7
Pandoraea_sp._RB-44	44	105.4
Streptococcus_suis_GZ1	26	105
Treponema_azotonutricium_ZAS-9	31	104.5
Escherichia_coli_O7::K1_str._CE10	11	102.6
Streptococcus_agalactiae_A909	66	102
Escherichia_coli_str._'clone_D_i14'	57	98.8
Escherichia_coli_BL21(DE3)	21	97.8
Flavobacteriaceae_bacterium_3519-10	92	95.1
Escherichia_coli_042	56	93.7
Bacillus_megaterium_WSH-002	101	92.5
Yersinia_pestis_D182038_Salmonella_enterica subsp._enterica_serovar	18	91.5
Choleraesuis_str._SC-B67	41	90.5

Table 2.1 continued

Desulfomicrobium_baculatum_DSM_4028	10	90
Cronobacter_sakazakii_SP291	27	89.2
Bacteroides_fragilis_NCTC_9343	1	82.9

As an additional benefit, lighGBM assigned an importance score to each strain by estimating the increase in error rate caused by removing that strain from the set of predictors (Figure 2.2c, Table 2.1). This brought us a chance to further reduce the size of biomarkers and optimize the discrimination ability. We sorting the 40 strains with regard to the importance score, and successively added 5 strains in a step into the training set of model training. We found that the model with 15 strains achieved the best performance, with the highest AUC of 79.55% (Figure 2.2b). Case by case model training and test further validated these 15 strains as signature strains (Figure 2.3 and Table 2.2).

Table 2.2: The 15 core strains

Core strains
D.orientis DSM 765
D.propionicus DSM 2032
Carnobacterium sp. 17-4
T.roseus DSM 18391
B.thuringiensis YBT-1518
B.gladioli BSR3
Nitrosomonas sp. AL212
Eggerthella sp. YY7918
Coprococcus catus GD-7
S.pneumoniae SPN034156
L.pneumophila 2300 99 Alcoy
P.putida NBRC 14164
M.nodulans ORS 2060
B.aphidicola str._Ak (Acyrthosiphon kondoi)
M.smegmatis str._MC2 155

2.2.6 Species-species co-occurrence networks

With the relative abundance of taxonomic composition across a group of samples, we derived the correlation of all genera, species and strains with Spearman correlation coefficient. Only Spearman correlations with $P < 0.05$ were accepted. Otherwise, the correlation coefficients (R) were manually set to be zero. R distributed in the range [-1,1]. In order to systematically summarize the co-occurrence network pattern, we counted the number of feature pairs with R in the corresponding intervals as shown in Figure 2.9a, b and c. The number was scaled and Gaussian function was found to be well fit the bell-shape curve. The pattern transformation of co-occurrence network was obvious on all the three taxonomic ranks (i.e., genus, species, strain).

2.2.7 Species interaction model construction and performance evaluation

We designed a four-step pattern recognition procedure to identify cases by species interaction networks. Firstly, we constructed a graph with all links representing species-species validated correlation ($P < 0.05$). The most intensively connected species was selected from the graph as the determinative feature for specimen selection in the current iteration. Basing on the relative abundance distribution of this selected feature across samples, mean value was adopted to partition the population into two separate groups, i.e., the high abundance group and the low abundance group. With each group as a specimen, the correlation between any two features was calculated as described above, resulting in a R (correlation coefficient) matrix and a corresponding P value matrix. The valid values ($P < 0.05$) in R matrix were classified into 20 bins as one observation. The selected feature and its strong connected features ($abs(R) > 0.6$) were removed from the graph and the remaining features were iterated into the next loop for feature selection and sampling. After 100 iterations for both controls and cases, there were 400 specimens in total. A neural network classifier with 10 hidden neurons was constructed for cases discrimination, the performance of which was assessed by receiver operating characteristic analysis. The AUC and corre-

sponding 95% confidence intervals for training data sets (1000 bootstrap replicates) were 99.63% ($98.38\% \sim 99.97\%$).

2.2.8 Index deduction for coordinate network

We have elucidated the complex interactions among a list of features. The network was evaluated by several indices as following, with reference[114]. Take COG category as one examples, which was the most complex case. Supposing that a network is supported by 26 COG categories, namely, $\{X_m\}_{m=1}^{26} = \{A, B, ..., Z\}$, and $G(X_m) = \{x_{mi}\}_1^{g(X_m)}$ is a set of COGs involved in X_m, where $g(X_m) = |G(X_m)|$ is the cardinality of $G(X_m)$. For any given prontein $x_{mi} \in G(X_m)$ corresponding to a node in the network, the existence of association between x_{mi} and another protein $x_{nj} \in G(X_n)$ is given by

$$\delta(x_{mi}, x_{nj}) = \begin{cases} 1, \text{if } |r(x_{mi}, x_{nj})| \geq 0.6 \text{ and } p(x_{mi}, x_{nj}) \leq 0.05; \\ 0, \text{otherwise,} \end{cases} \tag{2.1}$$

where R and P showed the correlation coefficient and corresponding P-value between the two COG x_{mi} and x_{nj}. The associated COG set of x_{mi} is defined as the neighbor node set:

$$A(x_{mi}) = \{x_{nj} x_{nj} \neq x_{mi}; \delta(x_{mi}, x_{nj}) = 1; x_{nj} \in G(X_n)\}, \tag{2.2}$$

where $n = 1, 2, ..., 26$. Let $a(x_{mi}) = A(x_{mi})$ to be the count of associated COGs of x_{mi}, and

$$b(x_{mi}) = \{(x_{m_1 i_1}, x_{m_2 i_2})\} \tag{2.3}$$

$$(x_{m_1 i_1}, x_{m_2 i_2} \in A(x_{mi}); \delta(x_{m_1 i_1}, x_{m_2 i_2}) = 1; i_1 \neq i_2) \tag{2.4}$$

to be the count of associations among COGs in COG category $A\,(x_{mi})$. The active index of COG x_{mi} is thus evaluated by a modified definition of cluster coefficient:

$$c(x_{mi}) = \frac{a(x_{mi})(a(x_{mi}) - 1) + 1}{2b(x_{mi}) + 1} - 1. \tag{2.5}$$

Another two indices for evaluation of COG categories were also worth attention. To measure the internal interaction of category X_m, let

$$v(x_{mi}|X_m) = \sum_{\substack{k=1 \\ k \neq i}}^{g(X_m)} \delta(x_{mi}, x_{mk}),\, x_{mi} \in X_m \tag{2.6}$$

denote the vertex degree of COG in X_m, then the internal complexity of association among COGs in X_m is measured by

$$\sum_{i=1}^{g(X_m)} v(x_{mi}|X_m) \ln v(x_{mi}|X_m), \tag{2.7}$$

as the definition of local network comlexity. For further comparision of different COG categories, the index got normalized by its maximum values as

$$t(X_m) = \frac{\sum_{i=1}^{g(X_m)} v(x_{mi}|X_m) \ln v(x_{mi}|X_m)}{g(X_m)(g(X_m) - 1) \ln(g(X_m) - 1)}. \tag{2.8}$$

Similarly consider the interaction between two COG categories X_m and X_n, and let

$$\begin{aligned} v(x_{mi}|X_n) &= \sum_{j=1}^{g(X_n)} \delta(x_{mi}, x_{nj}), \\ v(x_{nj}|X_m) &= \sum_{i=1}^{g(X_m)} \delta(x_{mi}, x_{nj}), \end{aligned} \qquad x_{mi} \in X_m, x_{nj} \in X_n. \tag{2.9}$$

2.3 Results

2.3.1 Overview of gut microbiome in each sample using FTU

In this study, we identify FTUs at a resolution considering functional orthologous as genes classified into the same COG[106, 107] and taxonomical unit of genes aligned onto the same reference genomes of one strain in NCBI[105, 109](see details in Data and methods). As a result, we identified $747,687$ FTUs in total, which consists of $6,253,507$ non-redundant gene sequences covering 33.65% of the DNA reads. Based on this combined profiling, we then deduced the relative abundance of FTUs across $1,659$ samples, subsequently obtained the FTU relative abundance matrix, which was the foundation of the follow-up analysis.

To investigate the genetic landscape of gut microbiota from both functional and taxonomic perspectives, we computed the cumulative number of FTUs present in any combination of n sample groups (with $n = 1-6$), demanding the FTUs are with relative frequencies higher than 10^{-8} to eliminate random mapping bias (Figure 2.1b). Here, we recognized the FTUs in a set of samples as an FTU pool. Integrating all samples in the six disease groups, we obtained an entire FTU pool of $747,687$ FTUs, with the case pool of $729,375$ FTUs and the control pool of $691,419$ FTUs. The majority ($673,107$ FTUs, 90.0% of the entire FTU pool) of the FTUs are common in both the case and the control FTU pools. However, comparing to the FTU pool of the controls with $18,312$ (2.6%) control-only FTUs, the one of the cases showed more expansive, with $56,268$ (7.7%) case-only FTUs, reflecting the higher divergence of the gut microbiota from the cases. As shown in Figure 2.1b, the FTU pools kept growing along with sample groups added in for both cases and healthy controls, suggesting the divergence of gut microbiome has not yet saturated in cases with different diseases as well as the healthy individuals even in such a large scale study. Nevertheless, it is obvious that the curves tended to converge with more sample groups added in.

To further study the characteristics of the human gut microbiome on the basis of F-

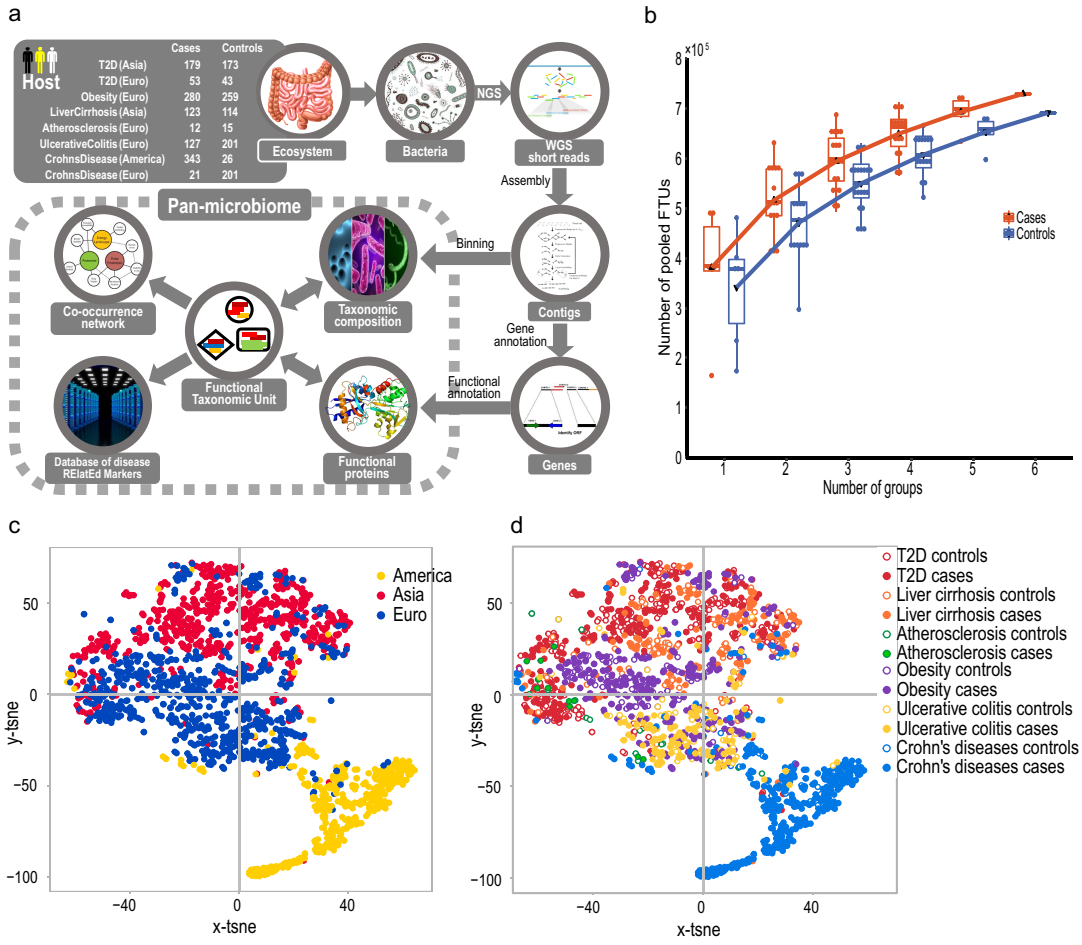

Figure 2.1: Overview of gut microbiome in each sample using FTU. **a**, Schematic diagram of the data processing workflow (details in Data and methods). **b**, Number of FTUs captured by number of investigated sample groups. For n sample groups ($n = 1, 2, ..., 6$), the number of FTUs presented in all possible group combinations (which is $6!/\,[n!(6-n)!]$) were separately measured and statistically shown by box and whisker plots for both cases (tangerine) and controls (blue). Boxes denote the interquartile range (IQR) between the first and third quartiles and the line inside denotes the median. The $\triangle$ and $\bigtriangledown$ denote the mean values of the cases and the controls, respectively. Whiskers denote the lowest and highest values within 1.5 times IQR from the first and third quartiles, respectively. Circles denote outliers beyond the whiskers. **c** and **d**, Reduced dimensional representation of samples in different groups performed by t-SNE algorithm. The grouping criteria were shown in different colors and styles on the legends. **c**, Samples were grouped by ethic background regardless of pathological conditions. **d**, Samples were grouped by pathological conditions irrespective of ethic background.

TU pool, we computed the distances between samples according to the relative abundance of the pooled FTUs. The t-SNE technique was then performed to obtain the reduced dimensional representation of all investigated samples. In order to demonstrate the sample clustering patterns, two grouping criteria were applied, the ethic/racial background and the pathological conditions, as shown in Figure 2.1c and 2.1d, respectively. Samples with consistent ethnic/racial background (i.e. European, American and Asian) were clearly observed to gather together, revealing that ethnic/racial background was one of the strongest associations of microbial community members (Figure 2.1c), which supported the previous finding of Tanya *et. al*[115]. Samples of intestinal microbial dysbiosis resulted from the same pathological disorders exhibited similar result, indicating different pathological disorders dragged the microbial communities into different directions (Figure 2.1d). This finding motivated us to integrate samples with same pathological disorder but from multiple geographical regions as one group for further comparative analysis, which was expected to elucidate the universal biological mechanisms by eliminating the deviations from inconsistent racial backgrounds.

2.3.2 The core FTUs of all samples and the signature strains

The exploration of the common set of FTUs is of equal importance as the FTU pool. For this purpose, we recognized the FTUs shared by more than 90% of the samples within a pathological group as shared FTUs. Applying similar approach as the construction of FTU pool, we computed the number of shared FTUs commonly presented in all combinations of n sample groups and illustrated as shown in Figure 2.2a. With sample groups sequentially added in, both curves exhibited a trend of convergence, of which the one of the controls is more noticeable. The common sets of core FTUs with $1,095$ and $12,494$ for cases and controls, respectively, were finally obtained by including all sample groups of six investigated pathological conditions. Comparing to the cases, the healthy controls possessed a much more inclusive common set of core FTUs but a relatively more compact FTU pool

at all combinations of group scales (2.1b and 2.2a). This showed that the gut microbial community of controls are more similar to each other than those of the cases, which are more differentiated in their own way, suggesting the different directions of alterations for microbiota under various pathological conditions.

By tracing back the taxonomic information of the core FTUs, we obtained 40 strains shared by all cases and 255 ones by the controls. The 40 shared strains of the cases were found to be universally presented among all samples of both cases and controls, with ranking on the top 100 abundant taxa (Table 2.1). We then regarded these 40 strains as the core strains. To test whether the core strains can be used to identify cases from controls, we trained a model of lightGBM (highly efficient boosting decision tree)[91] in a training set of the cases and controls using the profiles of these 40 strains. We evaluated the model performance with a ten-fold cross-validation approach and scored the discrimination ability by a receiver operating characteristic (ROC) analysis. The predictive power of model was assessed by the AUC (area under the ROC curve), which was 78.32%.

To further determine the representativeness of these 40 strains, we investigated the species-species co-occurrence network. Only connection bounds with absolute values of Spearman correlation larger than 0.5 ($P < 0.05$) are showed in Figure 2.2d and Figure 2.2e. We found that, compared to the controls, more positive and less negative interaction pairs were observed in the cases. This distinct pattern transformation of the co-occurrence network based on the core strains due to the difference of host pathological conditions was consistent with later investigations on global species co-occurrence network. Therefore, it validated that these 40 strains can be served as core and representative taxa of the gut ecological community.

We carried out more experiments to further simplify the set of signature strains whilst achieving better prediction performance. According to the importance score of each strain assigned by lighGBM, we sequentially selected a number of strains in a stepwise manner, with one strains added in at each step, as the model features and evaluated the behavior

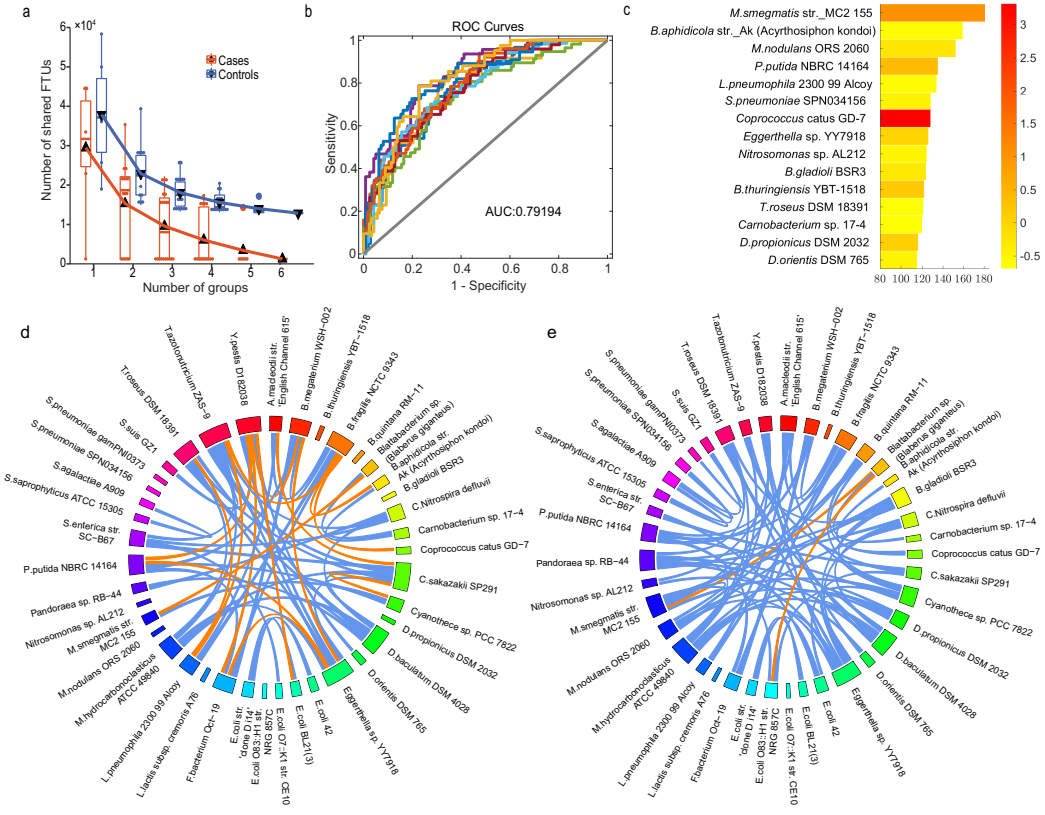

Figure 2.2: The common set of FTUs and the signature strains. **a**, Number of shared FTUs plotted as a function of number of sample groups investigated. The measurement of shared FTUs commonly presented in all combinations of n sample groups and its illustration were similar to Figure 2.1b. The cases and controls were displayed in tangerine and blue respectively. **b**, Performance evaluation of the lightGBM-based discrimination model trained on the 15 signature strains. The predictive power was scored by ROC analysis with a tenfold cross-validation approach. The average AUC reached 79.55%, suggesting the abundance of the signature strains can be employed as powerful biomarkers. Additionally, the model assigned an important score to each strain to measure its contribution to the discrimination ability as shown in **c**. The bar lengths indicate the importance of the strains, and colors represent their average relative abundances. **d**, Co-occurrence network of the signature strains distributing among the controls. **e**, Co-occurrence network of the signature strains distributing among the cases. In both **d** and **e**, the orange ribbons represent negative correlations, whereas the blue ones represent positive connections. The length of chords around the periphery is proportional to the number of connections.

of models with AUC. We found that as the feature set added up to 15 strains, the model performance didn't improve any more (Figure S1). In fact, with these 15 strains (Table 2.2), the discrimination model achieved better performance than the former model built with all the 40 strains. As shown in Figure 2.2b, cases were correctly identified with an average AUC of 79.55%, suggesting that these signature strains can be employed as universal signature taxa for different diseases. In addition, we validated the predictive ability of these signature strains as they could consistently predict host phenotype for different illnesses with high AUCs (see details in Figure 2.3). The average feature importance score of the ten-fold cross validation training models indicated several strains as the most discriminant for case identification. *Mycobacterium smegmatis* str._MC2_155, *Buchnera aphidicola* str._Ak_(Acyrthosiphon_kondoi) and *Methylobacterium nodulans* ORS_2060 were the three most discriminative bacteria in the models, with the first one had the highest discrimination score. We ascribed this result to the uneven distribution of these strains between cases and controls, also the relatively high abundance of these strains (Figure 2.2c). In summary, we demonstrated and recommended that the 15 signature strains can be considered as potential clinical biomarkers to indicate changes of the intestinal ecosystem.

2.3.3 Phylogenic and functional characteristics of the gut microbiome in diverse diseases

Having the FTU abundance matrix available, the profiles of taxonomic composition and functional constituent can be easily deduced. The result showed that the gut microbiota compositions are heterogeneous (Figure 2.4a), whereas the functional capabilities of these microbes are homogenous among different individuals (Figure 2.4b), which was consistent with previous report[17]. Regarding the microbial compositions, *Bacteroides*, *Escherichia* and *Coprococcus* were found to be the most abundant genera by averaging among all samples. The disease-specific genera species and strains of various sample groups or groups combinations can be deduced ($Q < 0.01$ in Wilcoxon rank-sum test). Case-control comparative analysis revealed that the relative distribution of hundreds of genera in samples

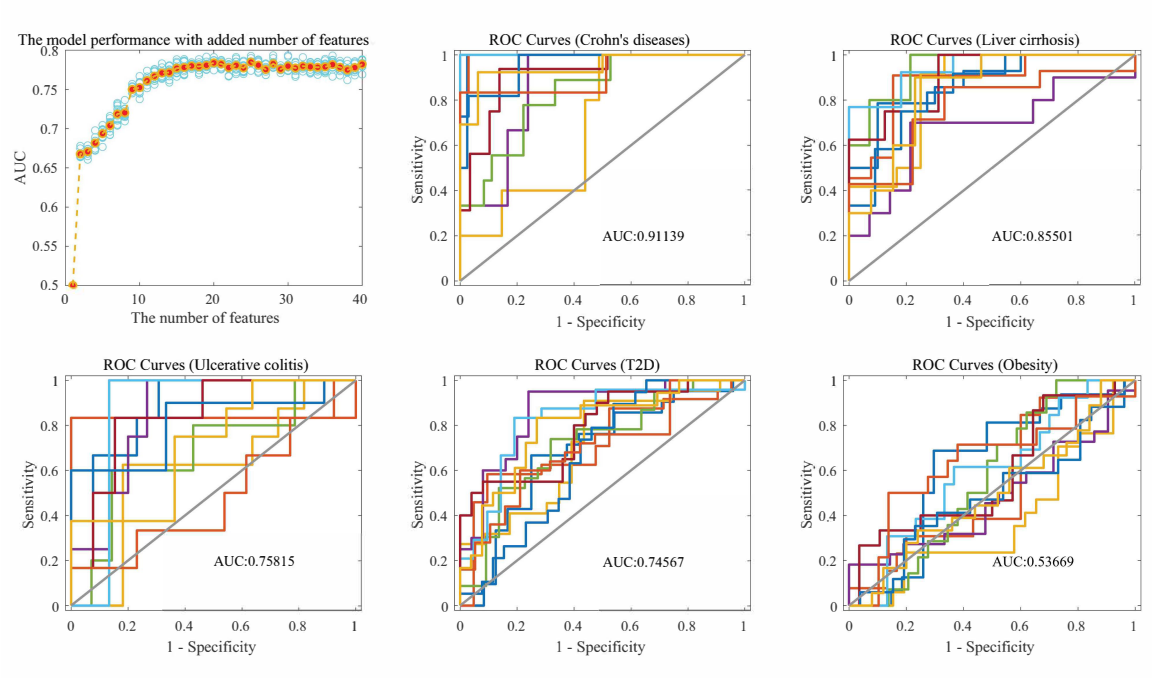

Figure 2.3: Performance evaluation of the lightGBM-based discrimination model trained on the signature strains.

with Crohn's disease, ulcerative colitis and liver cirrhosis were significantly different from their corresponding controls.It is worth noting that *Listeria*, which has been reported to be frequently associated with diarrhea and inflammatory response[116], was found to be significantly enriched in four types of diseases (i.e., Crohn's disease, ulcerative colitis, liver cirrhosis and obesity) comparing to their controls ($Q < 0.01$ in Wilcoxon rank-sum test), indicating that its enrichment plays a critical role in the development of multiple diseases. It is also worth mentioning that apparent transitions were easily observed within disease groups with multiple geographical origins, especially for taxonomic compositions. This finding further backed the conclusion obtained from Figure 2.1c and 2.1d.

With regard to the fact that the bacteria share genetic material, we should not only address the gut bacterial community by its members, but also consider the community as a whole and characterize it at the functional level of genes. As shown in Figure 2.4b, the dominant COG categories were replication, recombination and repair (L, 9.1% in health

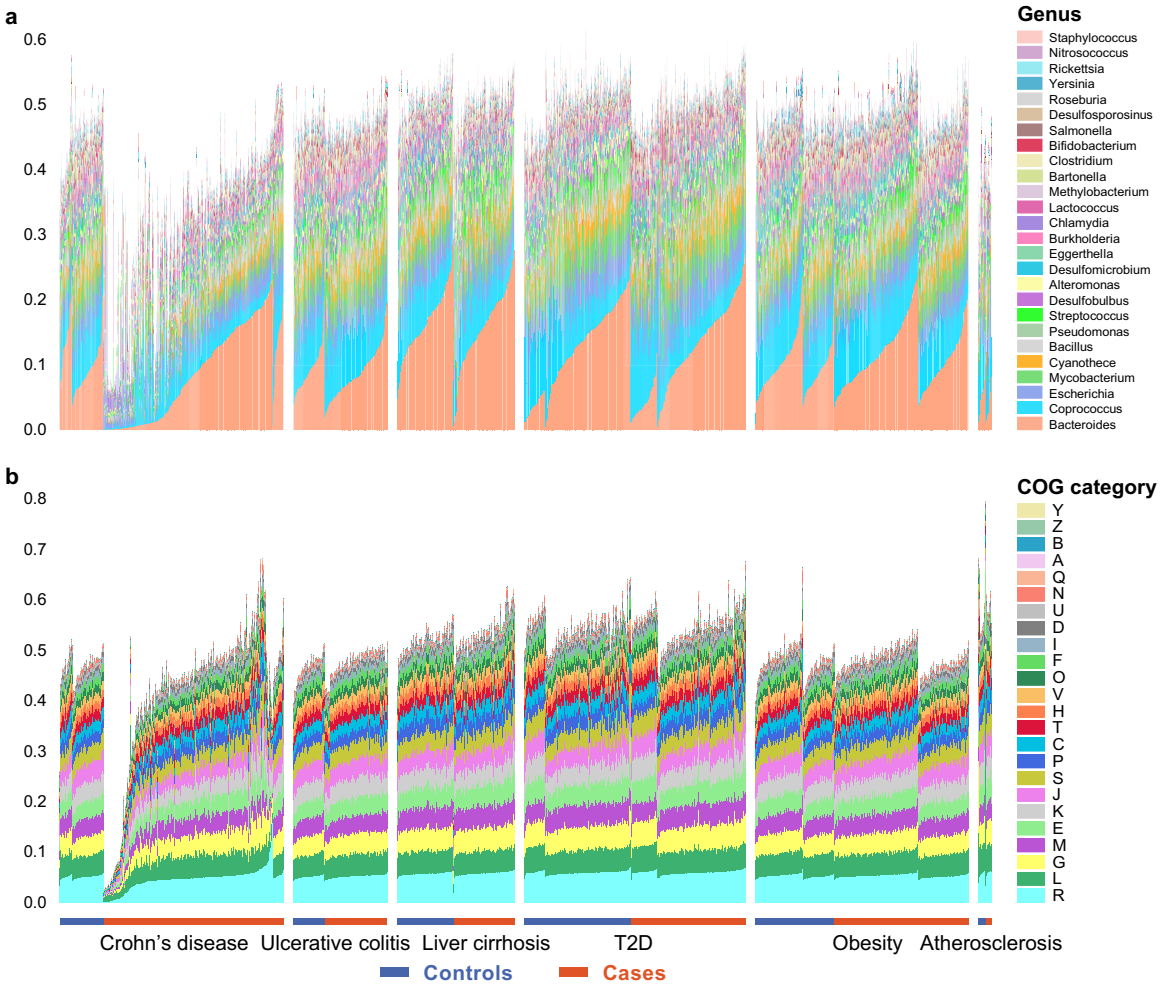

Figure 2.4: Abundance profiles of the genera and the COG categories across all samples. Vertical bars with various colors labeled on the right indicate the relative abundance of the microbial genera (**a**) and the COG categories (**b**), with blank representing 'others'. Samples are listed horizontally in the same order for both a and b as shown at the bottom, with blue and orange denote controls and cases, respectively. In both **a** and **b**, the legends are sorted by their average abundances among all samples, from the most at bottom to the least at the top.

controls and 8.9% in cases), carbohydrate transport and metabolism (G, 8.6% in health controls and 8.5% in cases), cell wall/membrane/envelope biogenesis (M, 7.6% in health controls and 7.7% in cases), amino acid transport and metabolism (E, 7.6% in health controls and 7.5% in cases), transcription (K, 7.5% both in health controls and cases) and translation, ribosomal structure and biogenesis (J, 7.1% in health controls and 7.2% in cases), which were important for basic life activities. However, when looking into specific functional orthologue groups (COG), highly diverse abundance distribution from sample to sample was observed. Functional case-control comparative analysis revealed that there existed a list of COGs ($Q < 0.01$ in Wilcoxon-rank sum test) significantly related to different diseases with more or less overlapping (Figure 2.5). Remarkably, Corhn's disease, ulcerative colitis and liver cirrhosis all led to distinct abundance distributions of a large number of COGs between cases and controls, implying that the three diseases were highly related to the transformation of the gut microbial metabolic functions, especially compared to the other two metabolic diseases T2D and obesity. Additionally, we noticed that homologs of the UspA protein (COG0589) enriched significantly in all five groups of cases in comparison with their controls. As a member of the universal stress proteins (USP), UspA constitute a natural biological defense mechanism which could help the organism surviving through nutrient starvation, the presence of oxidants or other stress agents[117]. Studies discovered that UspA is playing a role in the invasion and virulence of some pathogens further supported our finding[118, 119]. We traced the taxonomic origination of COG0589 by assembly all related FTUs and listed the result in Table 2.3. It suggested most of the related taxa were disease-enriched. In addition, CelA, phosphotransferase system (PTS) cellobiose-specific component IIB was found to be enriched in Crohn's disease, ulcerative colitis, liver cirrhosis and obesity. The evidence that PTS acted as virulence regulation for several pathogens backed this discovery[120].

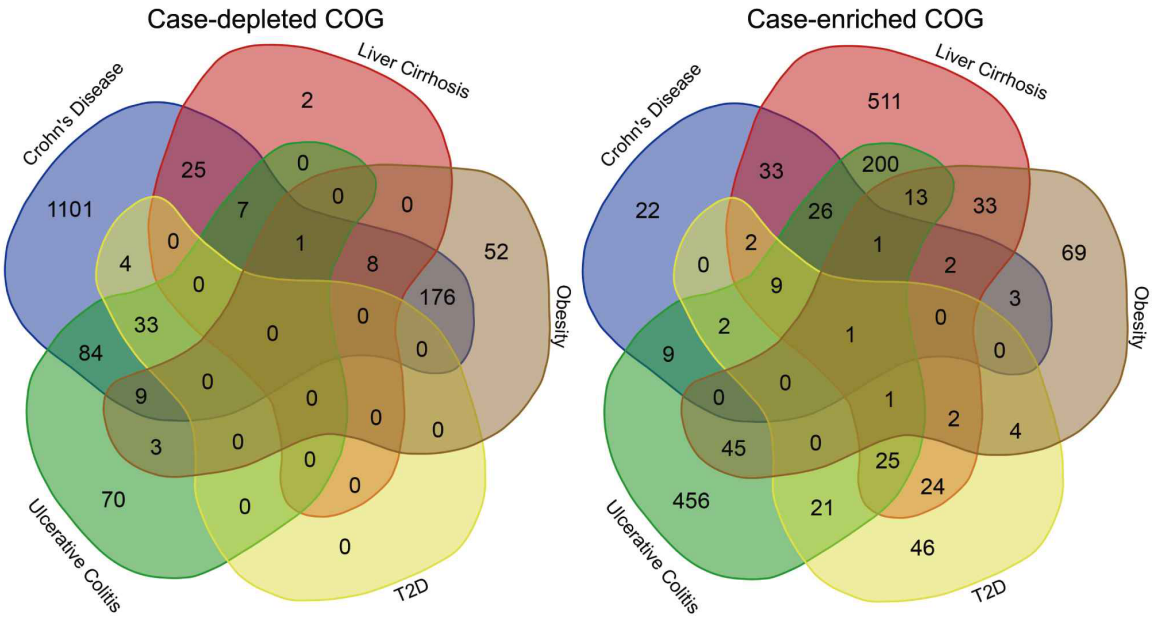

Figure 2.5: Venn diagram of the significantly disease-related COGs, which were revealed using case-control comparative analysis (Q ¡ 0.01 in Wilcoxon-rank sum test).

Table 2.3: The taxomnomic origination of genes in COG0589

Species	Case enriched	Case depelted
Bacteroides_fragilis	Yes	No
Carnobacterium_sp._17-4	No	No
Bifidobacterium_adolescentis	Yes	Yes
Cronobacter_sakazakii	No	No
Mycobacterium_smegmatis	Yes	Yes
Coprococcus_catus	Yes	No
Escherichia_coli	Yes	No
Synechococcus_sp._CC9902	No	No
Bacillus_amyloliquefaciens	Yes	Yes
Cyanothece_sp._PCC_7822	No	No

Table 2.3 continued

Desulfomicrobium_baculatum	Yes	No
Eggerthella_sp._YY7918	No	No
Salmonella_enterica	Yes	No
Butyrivibrio_fibrisolvens	No	No
Lactococcus_lactis	Yes	No
Bartonella_clarridgeiae	Yes	No
Pseudomonas_putida	No	No
Legionella_pneumophila	No	No
Chlamydia_trachomatis	Yes	No
Burkholderia_gladioli	Yes	No
Treponema_azotonutricium	No	Yes
Lactobacillus_casei	Yes	Yes
Streptococcus_pneumoniae	Yes	No
Pandoraea_sp._RB-44	No	No
Mesorhizobium_loti	Yes	No

2.3.4 Co-occurrence network pattern characterization of the gut microbiome

Although on the basis of abundance profiling, common markers of microbial community members and functional elements in various sample groups have been dug out, the underlying mechanism of how the microbial community is associated with diseases has yet to be elucidated thoroughly. Due to the inherent complexity of the gut microbial community as being part of natural ecosystems, studying each organism in isolation is far from enough. Co-occurrence patterns, as described above, are able to show how particular organisms in a system occur together and vary with the host pathological conditions. Considering the microbial community as a whole, we analyzed both the taxonomic and functional co-

occurrence network to characterize the gut ecosystem more systematically.

Firstly, we discovered that the distinct patterns of species correlation network in cases and controls elucidate the community-level adaptive mechanism. The gut microbiota is an ecosystem with many biological interactions. Together with host effect, dietary habits, antibiotics and other external factors, interactions between microbes have been revealed to be good implications in community assembly[121]. Relationships analogous to macro-ecological 'checkerboard pattern' of organismal co-occurrence have also been observed in microbial community due to competition and cooperation[122, 123]. With metagenomic data, species-species co-occurrence network analyses have provided a new dimension in studies of symbiotic microbial communities.

In this study, as a large number of samples covering six types of diseases and their healthy controls were investigated, we were able to conduct a comprehensive case-control comparative analysis on the interacting species pairs. We found that the positive interaction potentials of the cases are significantly higher than those of the controls. On the contrary, more negative interaction pairs were observed in controls than cases. This phenomenon can be observed when conducting comparative analysis on different phylogenic classification levels (i.e., genus, species and strain, as well as within pathological groups, see detail in Figure 2.9a, 2.9b and 2.9c, and 2.6),and Table 2.4). The consistent pattern shift suggested there exists a potential community-level adaptive mechanism to suit the dysbiosis of micro-ecology in the gut of cases. As proposed in previous studies[123, 124], habitat filtering was the dominant structuring force in the gut microbiome,which lead to the conclusion that species with negative connections tended to be complementary pairs, whereas positive correlations implied competitions between each other, with regarding to relative abundance distribution across samples. Therefore, more microbial members in the gut of healthy controls and cases are involved in cooperation and competition with each other, respectively. This can also be clearly observed in the network diagrams for both species level and genus level (Figure 2.7 and Figure 2.8, respectively).

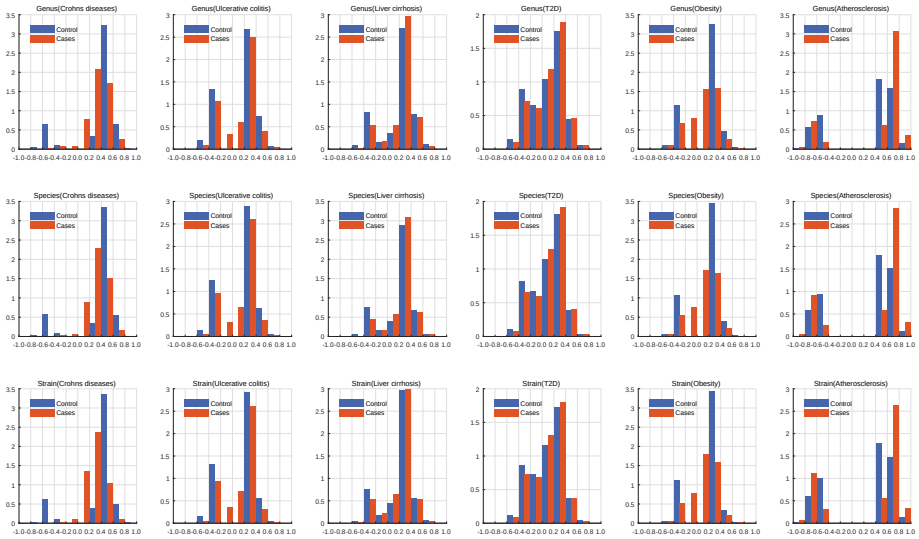

Figure 2.6: Distributions of correlation coefficient (R) of healthy controls and cases between genera (the first row), species (the second row) and strains (the third row).

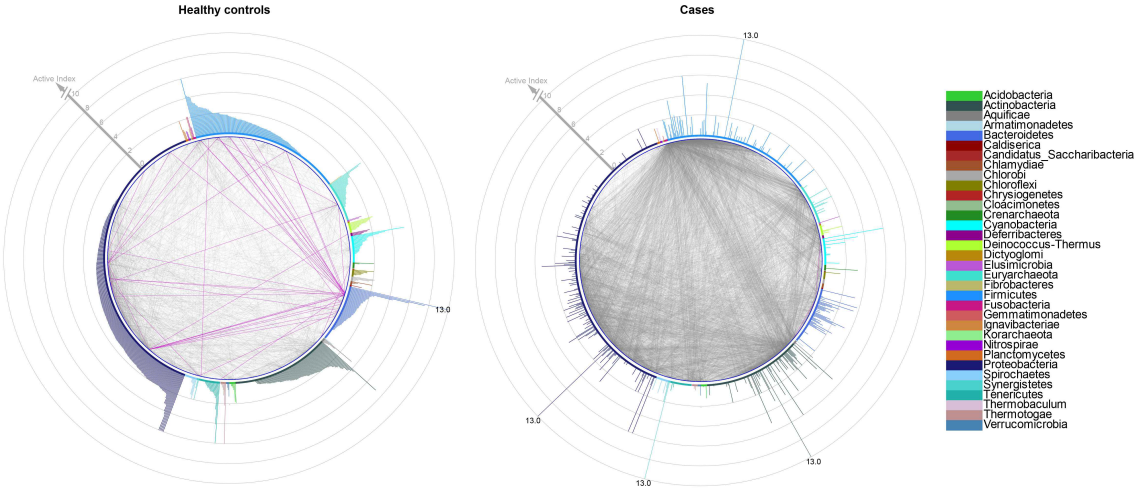

Figure 2.7: Species level taxonomic mutually dependent relationship. Interacted strength between pairwise species regarding their abundance distribution in controls and cases, respectively. The deeper the color, the stronger the interaction. Red and blue lines to represent negative and positive connection, respectively.

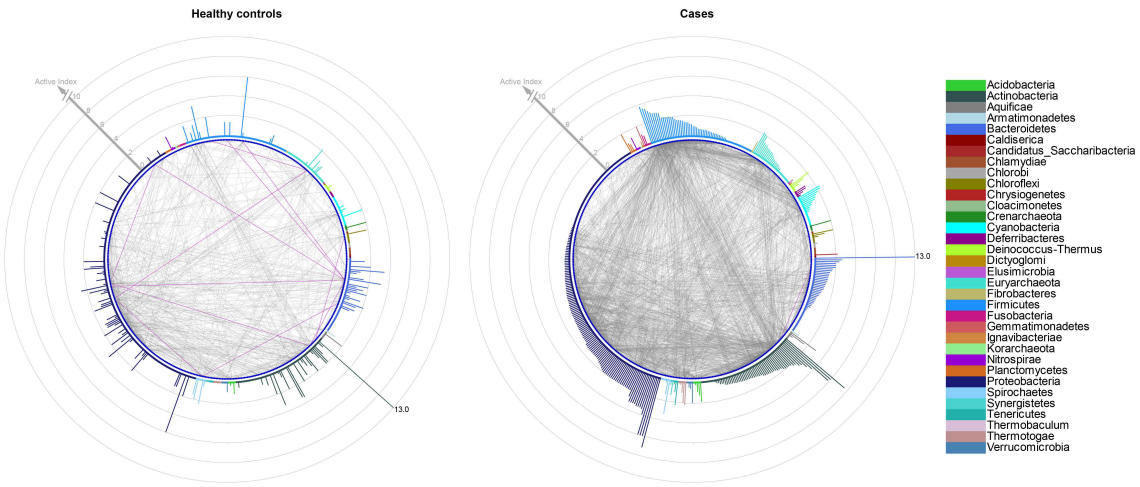

Figure 2.8: Same as Figure 2.7, This figure is on genus level

Table 2.4: Co-occurrence species pairs

Complementary pairs ($r < -0.6$, $P.value < 0.05$)		
	Healthy controls	Cases
CES-CES	47	20
CES-NCES	800	213
NCES-NCES	3542	412

Competetive pairs ($r > 0.6$, $P.value < 0.05$)		
	Healthy controls	Cases
CES-CES	29	594
CES-NCES	405	3061
NCES-NCES	1409	5029

CES: case-enriched species
NCES: not case-enriched species

In order to further validate this finding and distinguish the interaction patterns between healthy and cases, we employed a pattern recognition technique to identify cases from controls by species interaction networks. As shown in Figure 2.9d and described in Data and methods, a neural network discriminator was constructed with the training and tenfold cross-validation AUC achieving 0.9963 (within confidence interval 0.9838-0.9997, see Fig-

ure 2.9e) and 0.9428, respectively. It shed light on defining a healthy or unhealthy micro-ecology of human gut at a systematic level. We can apply this model into time-series samples or different set of samples to systematically characterize the co-occurrence pattern of one type of microbial ecosystem.

To explore the interactions between species and its underlying mechanism, we categorized the interactions into three types according to the relative richness on either side of the pair for further investigation. The interactions between both case-enriched species (CES-CES), case-enriched species and not case-enriched species (CES-NCES), and both not case-enriched species (NCES-NCES). When looking into those strongly correlated pairs ($|R| > 0.6, P < 0.05$), we found that the correlations of different interaction types were significantly varied with the status of pathological disorders (chi-square test, $P = 6\times10^{-16}$). Specifically, with regard to the competitive relationship, the positive correlations expanded extensively from healthy controls to cases, especially for the CES-CES pairs, which increased by over 20 times. On the contrary, the amount of cooperation pairs were dramatically decreased from healthy controls to cases, especially for the NCES-NCES pairs with a reduction of nearly tenfold. Integrating with the discoveries of forementioned studies[123, 124] and our current observations, it is implied that in a healthy host the microbial community members cooperated with each other in a harmonious and 'peaceful' manner. While during the gut microbiota alteration of the host from healthy to various pathological disorders, the CES, which are highly potential pathogens, dragged a large number of community members into a 'war' of largely increased competitions. They competed with other CES and the NCES to survive through the defective mechanism and further colonized the gut. As we known, the gut epithelium was protected by a layer of mucus composed of proteins known as mucins that are rich in fucose, galactose, sialic acid, N-acetylgalactosamine, N-acetylglucosamine and mannose. These sugars were harvested by saccharolytic members of the microbiota, such as *Bacteroidales* in the gut, which makes them available to species within the microbiota that lack of this capability. However, pathogenic bacteria in

the gut could also exploit the availability of these sugars to promote their own expansion. A strong evidence in our result is the interactions connected to one of the dominant species *Bacteroides fragilis*. In the co-occurrence network of healthy controls, it cooperated with 220 other species ($R < -0.4, P < 0.05$), most of which were NCES (Table 2.5 and Table 2.6). Nevertheless, in the network of the cases, all its complementary pairs disappeared with the emergence of 37 competing interactions instead. This has undoubtedly supported our speculation.

Table 2.5: Strongly correlated pairs in cases ($R > 0.6$ or $R < -0.4$, $P < 0.05$)

Genus1	Genus2	Correlation	Sign
Bacteroides_helcogenes	Bacteroides_fragilis	0.625	1
Bacteroides_salanitronis	Bacteroides_fragilis	0.603	1
Bacteroides_thetaiotaomicron	Bacteroides_fragilis	0.684	1
Bacteroides_vulgatus	Bacteroides_fragilis	0.679	1
Bacteroides_xylanisolvens	Bacteroides_fragilis	0.707	1
Ca.Azobacteroides_pseudotrichonymphae	Bacteroides_fragilis	0.732	1
Capnocytophaga_canimorsus	Bacteroides_fragilis	0.743	1
Capnocytophaga_ochracea	Bacteroides_fragilis	0.603	1
Chelativorans_sp._BNC1	Bacteroides_fragilis	0.657	1
Chitinophaga_pinensis	Bacteroides_fragilis	0.704	1
Chthonomonas_calidirosea	Bacteroides_fragilis	0.691	1
Desulfobacterium_autotrophicum	Bacteroides_fragilis	0.638	1
Dyadobacter_fermentans	Bacteroides_fragilis	0.663	1
Echinicola_vietnamensis	Bacteroides_fragilis	0.654	1
Emticicia_oligotrophica	Bacteroides_fragilis	0.652	1
Flavobacteriaceae_bacterium	Bacteroides_fragilis	0.707	1

Table 2.5 continued

Flavobacterium_johnsoniae	Bacteroides_fragilis	0.617	1
Leadbetterella_byssophila	Bacteroides_fragilis	0.694	1
Legionella_pneumophila	Bacteroides_fragilis	0.799	1
Maribacter_sp._HTCC2170	Bacteroides_fragilis	0.608	1
Niastella_koreensis	Bacteroides_fragilis	0.651	1
Odoribacter_splanchnicus	Bacteroides_fragilis	0.625	1
Paludibacter_propionicigenes	Bacteroides_fragilis	0.718	1
Parabacteroides_distasonis	Bacteroides_fragilis	0.752	1
Porphyromonas_gingivalis	Bacteroides_fragilis	0.637	1
Prevotella_denticola	Bacteroides_fragilis	0.618	1
Prevotella_intermedia	Bacteroides_fragilis	0.765	1
Prevotella_melaninogenica	Bacteroides_fragilis	0.616	1
Prevotella_sp._oral_taxon_299_str._F0039	Bacteroides_fragilis	0.617	1
Pseudomonas_putida	Bacteroides_fragilis	0.601	1
Solitalea_canadensis	Bacteroides_fragilis	0.637	1
Acinetobacter_baumannii	Bacteroides_fragilis	0.689	1
Alteromonas_macleodii	Bacteroides_fragilis	0.680	1
Anaeromyxobacter_sp._Fw109-5	Bacteroides_fragilis	0.624	1
Acinetobacter_baumannii	Bacteroides_fragilis	0.689	1
Alteromonas_macleodii	Bacteroides_fragilis	0.680	1
Anaeromyxobacter_sp._Fw109-5	Bacteroides_fragilis	0.624	1

Secondly, based on the functional profile of all samples, we integrated the COG abundance distribution of all samples and deduced the coordination network indices, i.e. internal complexity of individual COG categories and interacted strength between different COG

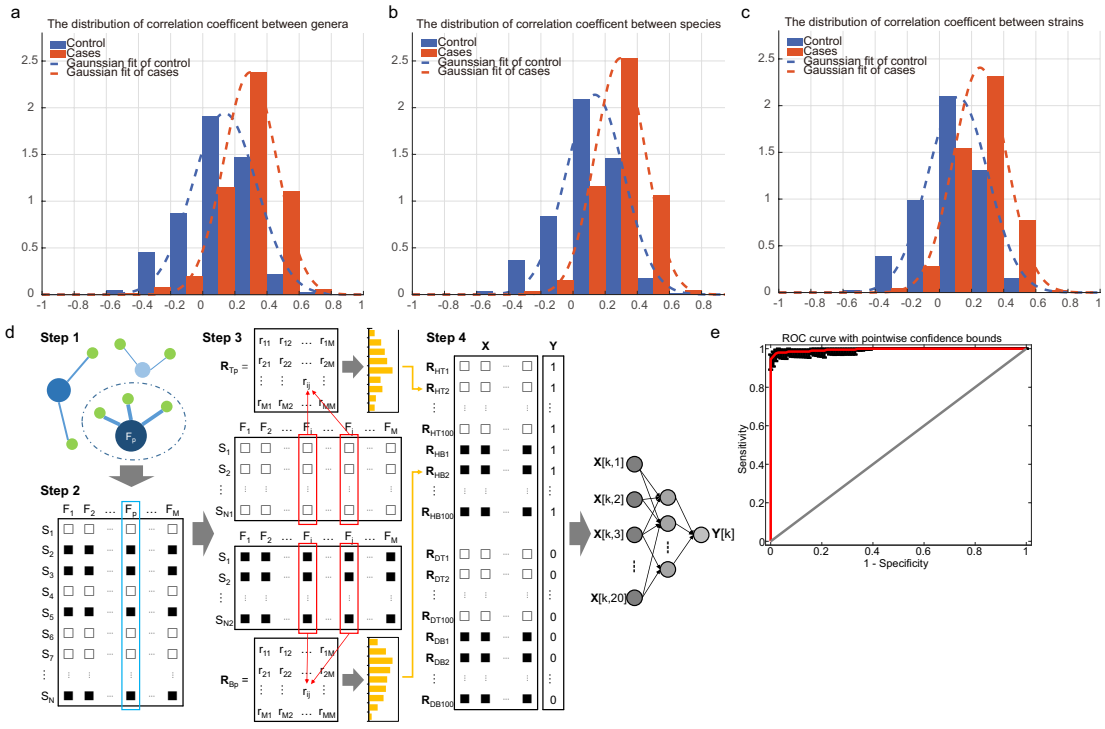

Figure 2.9: **a-c**, Distributions of correlation coefficient (R) of healthy controls and cases between genera (**a**), species (**b**) and strains (**c**). The height of the bars in the histograms represented the relative number of feature pairs with R's falling into corresponding intervals. The Spearman correlation should meet the criteria of $P < 0.05$, otherwise the relevant R's were manually set to be zero. The distributions were easily noticed to be of bell shapes that Gaussian distributions can be used for well fitting. **d**, Outline of the four-step pattern recognition procedure to identify cases based on the species interaction networks. Step 1, from the overall species -species correlation network, the most intensively connected species was selected as the determinate feature for specimen selection in current iteration. Step 2, based on relative abundance distribution of the selected feature across samples, mean value was adopted to conduct sampling as described in Data and methods. Step 3, for each specimen, the correlation between any two features was calculated, resulting in a R (correlation coefficient) matrix and a corresponding P value matrix. The valid values in each R matrix were classified into 20 bins as one observation. The selected feature in step 1 and its intensively connected features ($abs(R) > 0.6$) were removed from the network and the remaining features were iterated into the next loop of Step 1-3 for feature selection and sampling. Step 4, 400 specimens were obtained in total after 100 iterations for both controls and cases. Step 5, a neural network classifier with 10 hidden neurons was trained for cases discrimination, with the performance assessed by ROC analysis. The AUC and its 95% confidence intervals of training sets (1000 bootstrap replicates) were 99.63% and 98.38% - 99.97%, respectively (**e**).

categories (see Data and methods for detailed calculation). The result showed that the COG category N (cell motility) was highly internal complex (Figure 2.10a) although with relatively low abundance (Figure 2.4b), suggesting a prominent hallmark of cell motility. Further in the interacted network, cell motility (N) was intensively connected to signal transduction mechanisms (T) and cell cycle control/cell division/chromsome partitioning(D) (Figure 2.10b and 2.10c), especially for cases. This observation suggested that cell motility is important for bacterial colonization of the hosts, especially for pathogens, with regard to the fact that pathogens are more abundant in cases. Additionally, we detected that translation/ribosomal structure and biogenesis (J) linked closely to cell cycle control/cell division/chromsome partitioning (D), nucleotide transport and metabolism (F), transcription (K) and replication/recombination/repair (L), which implied that functional category J was the linkage center for these cellular activities.

Regarding the functional pathway profiles, we analyzed the coordination network of KEGG modules. Both networks of healthy controls (Figure 2.10d) and cases (Figure 2.10e) implied that KEGG module of flagellar assembly and bacterial chemotaxis were intensively connected with each other (Spearman correlation $R = 0.89$ for healthy controls and $R = 0.92$ for cases, $P < 0.05$). As both modules are crucial for the process of the host colonization[125], our finding implied that these two processes should be highly connected with each other. The higher level of connections for sample groups of cases indicated that the combination of these two modules is likely to help enhancing the establishment of a successful infection for pathogens. As flagellar assembly is the determinant factor for cell motility, this finding further backed the result of COG category internal complexity analysis. Carbon metabolism of healthy controls was found to be linked with other pathways more actively than that of cases, which can also be concluded from the higher active index of this pathway in networks (Figure 2.10b and 2.10c). It is suggested that bacteria of healthy controls conducted carbon metabolism more efficiently than which of the cases through intensively cooperating with relevant modules. In addition, we found that

lipopolysaccharide biosynthesis shown negative correlation with several metabolic pathways in healthy controls but not in cases. As we known, the lipopolysaccharide which constitutes the outer leaflet of the outer membrane for most Gram-negative bacteria is commonly referred to as an endotoxin[126]. This observation indicated the importance of trade off between different metabolic pathways to maintain a healthy microbial ecology.

2.3.5 A more comprehensive pan-microbiome revealed by members, functions and networks

Pan-microbiome has been proposed as the collection of all microbial members in a certain environment with stable microbial community[127, 128, 129], which have nonetheless provided important insights. Despite the currently limited knowledge regarding disease specific and graphical differences in human gut microbiology, the data in our paper allowed us to appreciate that the microbiomes of different cases consolidate to form a pan-microbiome pool that is larger than the microbiome of any single study. As a natural habitat for numerous microorganisms, human gut is a dynamic microbial ecological system. As the microbiota share genetic material, we should not only address the gut bacterial community by community members, but also should consider the community as a whole and characterize it at the functional level of genes. Here, with the availability of WGS data from human gut microbiota deep sequencing and FTUs, we can further explore the functions of the microbial members. In addition, we proposed the co-occurrence network analysis of community members and functional gene orthologs to characterize the human gut community on system level. All of these aspects contributed to a more comprehensive "pan-microbiome" and provided deep insights into how the gut microbiomes have diverged during the physiological condition shift.

2.4 Discussion

We have collected WGS data of gut microbiome from a wide range and large scale of hosts of diverse physiological conditions and various national backgrounds to obtain a global

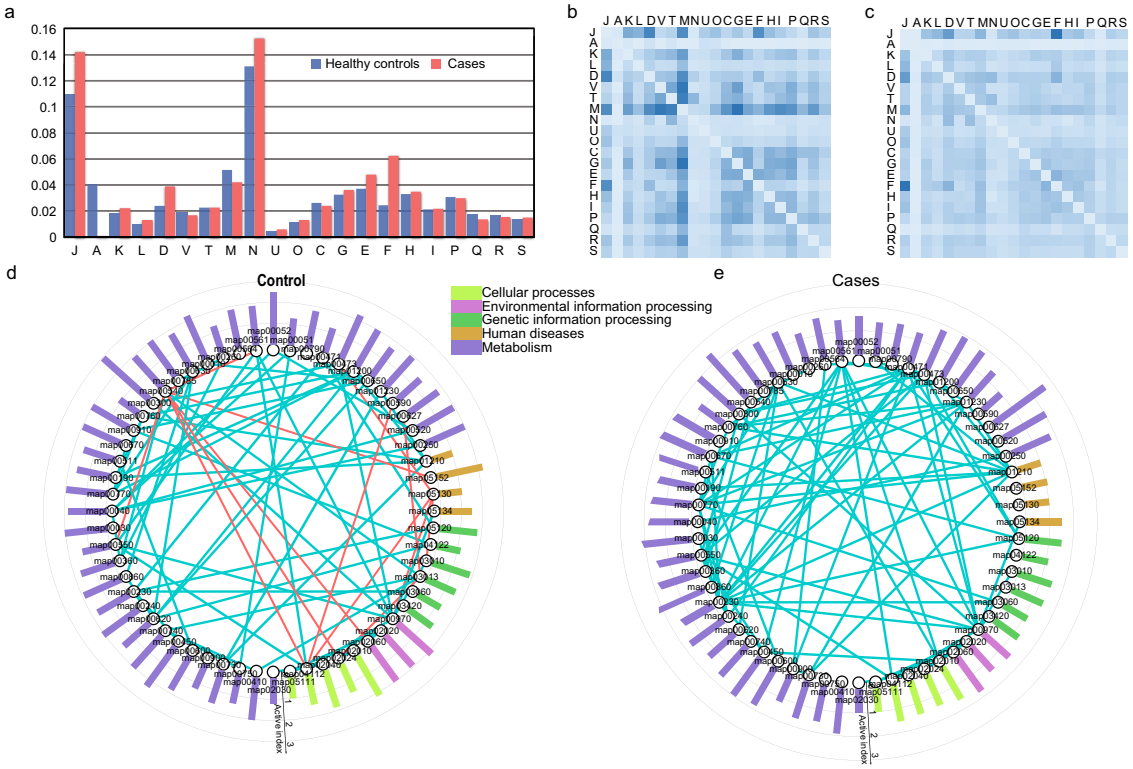

Figure 2.10: Characterization of the microbial community functions with co-occurrence network indices. **a**, Internal complexity of individual COG categories. **b** and **c**, Interacted strength between pairwise COG categories with regard to their abundance distribution in controls and cases, respectively. The deeper the color, the stronger the interaction. **d** and **e**, Coordination networks of KEGG modules for healthy controls and cases, respectively. Network active index of KEGG pathways were shown as the height of columns and strong correlations between different KEGG pathways were linked by red and blue lines to represent negative and positive connection, respectively.

framework of understanding to the dysbiosis of human gut microbial community. With the integration of both taxonomic classification and functional annotation, we organized a large number of sequencing reads into a pool of FTUs, the basic analysis units in this context microbial community characterization. All WGS short reads gathered from various sources were processed with a unified and well-designed workflow to make sure the analysis and results are comparable. Meanwhile, the relative abundance was normalized over all features for each sample to eliminate the variation caused by different sequencing depth. Further more, we adopted the Q value instead of P value as the criterion for determining disease specific features to keep criterion consistency among variant samples sizes of pathological groups.

The large collection of metagenomic data set from spanning hosts with standardized processing and analyzing workflow enabled us for a systematic investigation. With comparative analysis on FTU pools of all six sample groups, we identified 40 core strains and further handpicked 15 of them as signature strains, which optimized the separation of cases and controls. Case-control studies on genera compositions and COG profiles revealed that genus *Listeria* and protein homolog UspA (COG0589) were significantly more abundant in cases with various diseases than corresponding controls ($Q < 0.01$). Therefore, they are of highly potential to serve as universal biomarkers to aid therapeutic means aiming to amend gut ecological composition. The atherosclerosis group was excluded from Venn diagram (Table 2.2) drawing because of the variation insignificance between cases and controls, probably due to insufficient sample size.

The statistical result suggested that although taxonomic compositions varied from sample to sample, the functional compositions maintained consistent when classified into COG categories. This implied the functional stability of human gut microbiota at the community-level. Basic functions such as replication/recombination/repair, carbohydrate transport and metabolism were the most abundant ones. Nevertheless, abundance was not the only factor to determine the significance of a feature, especially when regarding human gut as a

complex microbial ecology. We then proposed a mathematical model to compute the internal complexity and interacted strength of COG categories, separately for quantifying the activeness of the functions and the strength of connections between functional categories. The result showed that in addition to the forementioned basic abundance dominant functions, cell motility was of high internal complexity and connected intensively with signal transduction mechanisms, especially for cases. Similar coordinate network analysis based on KEGG modules further backed this finding. Besides all of the above, we believe our mathematical model also provided a new approach to investigate the significance of a feature in a complex system.

The most exciting discovery in this study is that we revealed the potential adaptive mechanism of human gut microbial community. Correlation networks of microbial members showed that microbes tended to be negatively connected in healthy controls, whereas positively connected in cases. This implied that the microbial community of a healthy hosts is in a harmonious mode with more cooperation connections between members. Until the pathogens started to perturb the original state and drag lots of members into a 'war' mode with increasingly competition connections during gut microbiota alteration of the host suffering from pathological disorder. As human gut has been reported as a habitat-filtering microbial ecology, this finding supplied potential adaptive mechanism for nutritional integration. Although some detailed evidences found in the coordination network strongly supported our speculation, a mechanistic and comprehensive understanding of those observations remained elusive. More rigorously designed experiments are needed to further verify this interesting discovery.

Though we could not avoid the deviation from the inconsistent sequencing platforms and uncertainty during annotation, we designed a uniform analysis pipeline form the start of raw short reads to reduce the potential deviation. We encourage large scale collection of samples with different kinds of diseases all around the world and conduct data sequencing with identical sequencing technique, which will be better. Additionally, longitudinal mon-

itoring of the microbiome of human gut will also benefit the analysis regarding human gut microbial community as an ecological system. Nevertheless, our systematic analysis greatly enhanced our understanding of the pan-microbiome. With all findings described above, our knowledge to the associations of gut flora and host physiological disorders has widened and deepened. Moreover, increasing number of gut microbiome samples relevant to more human diseases, and the development of multi-omics such as proteomics and metabolomics, are hoping to contribute to expand our comprehension of human gut pan-microbiome.

Table 2.6: Strongly correlated pairs in healthy controls($R > 0.6$ or $R < -0.4$, $P < 0.05$)

Genus1	Genus2	Correlation	Sign
Bacteroides_vulgatus	Bacteroides_fragilis	0.620	1
Bacteroides_xylanisolvens	Bacteroides_fragilis	0.602	1
Brucella_canis	Bacteroides_fragilis	0.630	-1
Ca.Azobacteroides_pseudotrichonymphae	Bacteroides_fragilis	0.686	1
Ca.Phytoplasma_mali	Bacteroides_fragilis	0.631	-1
Capnocytophaga_canimorsus	Bacteroides_fragilis	0.708	1
Capnocytophaga_ochracea	Bacteroides_fragilis	0.655	1
Chitinophaga_pinensis	Bacteroides_fragilis	0.624	1
Dyadobacter_fermentans	Bacteroides_fragilis	0.707	1
Echinicola_vietnamensis	Bacteroides_fragilis	0.604	1
Flavobacterium_johnsoniae	Bacteroides_fragilis	0.612	1
Lactobacillus_salivarius	Bacteroides_fragilis	0.611	-1
Leadbetterella_byssophila	Bacteroides_fragilis	0.680	1
Legionella_pneumophila	Bacteroides_fragilis	0.843	1
Methanobrevibacter_smithii	Bacteroides_fragilis	0.606	-1
Paludibacter_propionicigenes	Bacteroides_fragilis	0.713	1

Table 2.6 continued

Parabacteroides_distasonis	Bacteroides_fragilis	0.674	1
Peptoclostridium_difficile	Bacteroides_fragilis	0.622	-1
Prevotella_intermedia	Bacteroides_fragilis	0.675	1
Pseudomonas_brassicacearum	Bacteroides_fragilis	0.621	-1
Solitalea_canadensis	Bacteroides_fragilis	0.618	1
Sulfurimonas_autotrophica	Bacteroides_fragilis	0.645	-1
Thermoanaerobacter_sp._X514	Bacteroides_fragilis	0.616	-1
Tistrella_mobilis	Bacteroides_fragilis	0.618	-1
Alteromonas_macleodii	Bacteroides_fragilis	0.613	1
Bacillus_halodurans	Bacteroides_fragilis	0.613	-1
Bacillus_licheniformis	Bacteroides_fragilis	0.664	-1
Bacillus_sp._JS	Bacteroides_fragilis	0.609	-1
Bacillus_subtilis	Bacteroides_fragilis	0.602	-1
Acidobacterium_capsulatum	Bacteroides_fragilis	0.575	-1
Acidovorax_avenae	Bacteroides_fragilis	0.455	-1
Agrobacterium_radiobacter	Bacteroides_fragilis	0.485	-1
Agrobacterium_vitis	Bacteroides_fragilis	0.565	-1
Alicycliphilus_denitrificans	Bacteroides_fragilis	0.586	-1
Alicyclobacillus_acidocaldarius	Bacteroides_fragilis	0.559	-1
Alkaliphilus_metalliredigens	Bacteroides_fragilis	0.478	-1
Amycolatopsis_orientalis	Bacteroides_fragilis	0.408	-1
Anaerobaculum_mobile	Bacteroides_fragilis	0.468	-1
Anaplasma_marginale	Bacteroides_fragilis	0.416	-1
Arthrobacter_chlorophenolicus	Bacteroides_fragilis	0.548	-1
Arthrobacter_phenanthrenivorans	Bacteroides_fragilis	0.439	-1
Azospirillum_brasilense	Bacteroides_fragilis	0.501	-1

Table 2.6 continued

Bacillus_amyloliquefaciens	Bacteroides_fragilis	0.507	-1
Bacillus_clausii	Bacteroides_fragilis	0.516	-1
Bacillus_coagulans	Bacteroides_fragilis	0.433	-1
Bacillus_cytotoxicus	Bacteroides_fragilis	0.569	-1
Bacillus_halodurans	Bacteroides_fragilis	0.613	-1
Bacillus_infantis	Bacteroides_fragilis	0.428	-1
Bacillus_licheniformis	Bacteroides_fragilis	0.664	-1
Bacillus_pseudofirmus	Bacteroides_fragilis	0.518	-1
Bacillus_pumilus	Bacteroides_fragilis	0.421	-1
Bacillus_sp._1NLA3E	Bacteroides_fragilis	0.592	-1
Bacillus_sp._JS	Bacteroides_fragilis	0.609	-1
Bacillus_subtilis	Bacteroides_fragilis	0.602	-1
Bacillus_toyonensis	Bacteroides_fragilis	0.476	-1
Bacillus_weihenstephanensis	Bacteroides_fragilis	0.411	-1
Bartonella_australis	Bacteroides_fragilis	0.442	-1
Bartonella_henselae	Bacteroides_fragilis	0.426	-1
Bdellovibrio_exovorus	Bacteroides_fragilis	0.476	-1
Beutenbergia_cavernae	Bacteroides_fragilis	0.457	-1
Bifidobacterium_adolescentis	Bacteroides_fragilis	0.524	-1
Bifidobacterium_animalis	Bacteroides_fragilis	0.591	-1
Bifidobacterium_bifidum	Bacteroides_fragilis	0.512	-1
Bifidobacterium_longum	Bacteroides_fragilis	0.458	-1
Borrelia_burgdorferi	Bacteroides_fragilis	0.401	-1
Borrelia_garinii	Bacteroides_fragilis	0.482	-1
Borrelia_recurrentis	Bacteroides_fragilis	0.435	-1
Bradyrhizobium_sp._BTAi1	Bacteroides_fragilis	0.440	-1

Table 2.6 continued

Bradyrhizobium_sp._ORS_278	Bacteroides_fragilis	0.466	-1
Bradyrhizobium_sp._S23321	Bacteroides_fragilis	0.409	-1
Brucella_canis	Bacteroides_fragilis	0.630	-1
Brucella_ceti	Bacteroides_fragilis	0.414	-1
Brucella_melitensis	Bacteroides_fragilis	0.534	-1
Buchnera_aphidicola	Bacteroides_fragilis	0.545	-1
Burkholderia_gladioli	Bacteroides_fragilis	0.447	-1
Burkholderia_multivorans	Bacteroides_fragilis	0.522	-1
Burkholderia_phenoliruptrix	Bacteroides_fragilis	0.400	-1
Burkholderia_phymatum	Bacteroides_fragilis	0.451	-1
Burkholderia_phytofirmans	Bacteroides_fragilis	0.503	-1
Burkholderia_sp._CCGE1003	Bacteroides_fragilis	0.523	-1
Burkholderia_sp._YI23	Bacteroides_fragilis	0.471	-1
Ca.Kinetoplastibacterium_crithidii	Bacteroides_fragilis	0.408	-1
Ca.Phytoplasma_mali	Bacteroides_fragilis	0.631	-1
Ca.Arthromitus_sp._SFB-mouse-Yit	Bacteroides_fragilis	0.517	-1
Cellulomonas_fimi	Bacteroides_fragilis	0.427	-1
Chlamydophila_pneumoniae	Bacteroides_fragilis	0.435	-1
Chlorobium_limicola	Bacteroides_fragilis	0.407	-1
Chromobacterium_violaceum	Bacteroides_fragilis	0.577	-1
Clostridium_difficile	Bacteroides_fragilis	0.405	-1
Clostridium_novyi	Bacteroides_fragilis	0.467	-1
Clostridium_perfringens	Bacteroides_fragilis	0.439	-1
Clostridium_tetani	Bacteroides_fragilis	0.586	-1
Conexibacter_woesei	Bacteroides_fragilis	0.460	-1
Coprococcus_catus	Bacteroides_fragilis	0.534	-1

Table 2.6 continued

Coprococcus_sp._ART55SLASH1	Bacteroides_fragilis	0.480	-1
Coriobacterium_glomerans	Bacteroides_fragilis	0.461	-1
Cronobacter_sakazakii	Bacteroides_fragilis	0.475	-1
Dehalobacter_sp._DCA	Bacteroides_fragilis	0.522	-1
Desulfatibacillum_alkenivorans	Bacteroides_fragilis	0.481	-1
Desulfomicrobium_baculatum	Bacteroides_fragilis	0.478	-1
Desulfotomaculum_carboxydivorans	Bacteroides_fragilis	0.463	-1
Desulfotomaculum_gibsoniae	Bacteroides_fragilis	0.482	-1
Desulfovibrio_salexigens	Bacteroides_fragilis	0.417	-1
Desulfurispirillum_indicum	Bacteroides_fragilis	0.559	-1
Desulfurococcus_mucosus	Bacteroides_fragilis	0.433	-1
Dickeya_dadantii	Bacteroides_fragilis	0.433	-1
Eggerthella_sp._YY7918	Bacteroides_fragilis	0.439	-1
Enterobacter_cloacae	Bacteroides_fragilis	0.444	-1
Ethanoligenens_harbinense	Bacteroides_fragilis	0.495	-1
Eubacterium_rectale	Bacteroides_fragilis	0.473	-1
Faecalibacterium_prausnitzii	Bacteroides_fragilis	0.496	-1
Flexistipes_sinusarabici	Bacteroides_fragilis	0.556	-1
Frankia_sp._EuI1c	Bacteroides_fragilis	0.412	-1
Geobacillus_sp._JF8	Bacteroides_fragilis	0.501	-1
Gloeobacter_kilaueensis	Bacteroides_fragilis	0.416	-1
Gordonibacter_pamelaeae	Bacteroides_fragilis	0.484	-1
Haemophilus_somnus	Bacteroides_fragilis	0.486	-1
Haliangium_ochraceum	Bacteroides_fragilis	0.432	-1
Haloarcula_hispanica	Bacteroides_fragilis	0.453	-1
Helicobacter_bizzozeronii	Bacteroides_fragilis	0.418	-1

Table 2.6 continued

Heliobacterium_modesticaldum	Bacteroides_fragilis	0.473	-1
Hyphomonas_neptunium	Bacteroides_fragilis	0.519	-1
Ketogulonicigenium_vulgare	Bacteroides_fragilis	0.504	-1
Kineococcus_radiotolerans	Bacteroides_fragilis	0.459	-1
Kocuria_rhizophila	Bacteroides_fragilis	0.545	-1
Kribbella_flavida	Bacteroides_fragilis	0.464	-1
Kytococcus_sedentarius	Bacteroides_fragilis	0.541	-1
Lachnoclostridium_phytofermentans	Bacteroides_fragilis	0.457	-1
Lactobacillus_brevis	Bacteroides_fragilis	0.460	-1
Lactobacillus_helveticus	Bacteroides_fragilis	0.557	-1
Lactobacillus_salivarius	Bacteroides_fragilis	0.611	-1
Lawsonia_intracellularis	Bacteroides_fragilis	0.449	-1
Leptospirillum_ferriphilum	Bacteroides_fragilis	0.405	-1
Leptotrichia_buccalis	Bacteroides_fragilis	0.504	-1
Leuconostoc_kimchii	Bacteroides_fragilis	0.523	-1
Listeria_monocytogenes	Bacteroides_fragilis	0.481	-1
Marinobacter_hydrocarbonoclasticus	Bacteroides_fragilis	0.426	-1
Melissococcus_plutonius	Bacteroides_fragilis	0.478	-1
Mesorhizobium_ciceri	Bacteroides_fragilis	0.453	-1
Mesorhizobium_opportunistum	Bacteroides_fragilis	0.481	-1
Methanobacterium_lacus	Bacteroides_fragilis	0.424	-1
Methanobrevibacter_ruminantium	Bacteroides_fragilis	0.473	-1
Methanobrevibacter_smithii	Bacteroides_fragilis	0.606	-1
Methanocaldococcus_fervens	Bacteroides_fragilis	0.457	-1
Methanocaldococcus_infernus	Bacteroides_fragilis	0.410	-1
Methanosaeta_harundinacea	Bacteroides_fragilis	0.499	-1

Table 2.6 continued

Methanosphaera_stadtmanae	Bacteroides_fragilis	0.431	-1
Methanothermobacter_marburgensis	Bacteroides_fragilis	0.456	-1
Methanothermobacter_thermautotrophicus	Bacteroides_fragilis	0.458	-1
Methylobacterium_extorquens	Bacteroides_fragilis	0.405	-1
Methylovorus_sp._MP688	Bacteroides_fragilis	0.413	-1
Modestobacter_marinus	Bacteroides_fragilis	0.539	-1
Mycoplasma_penetrans	Bacteroides_fragilis	0.438	-1
Myxococcus_stipitatus	Bacteroides_fragilis	0.449	-1
Myxococcus_xanthus	Bacteroides_fragilis	0.435	-1
Natranaerobius_thermophilus	Bacteroides_fragilis	0.455	-1
Natronomonas_pharaonis	Bacteroides_fragilis	0.462	-1
Neisseria_gonorrhoeae	Bacteroides_fragilis	0.441	-1
Nocardia_farcinica	Bacteroides_fragilis	0.565	-1
Nocardioides_sp._JS614	Bacteroides_fragilis	0.528	-1
Nocardiopsis_dassonvillei	Bacteroides_fragilis	0.405	-1
Novosphingobium_sp._PP1Y	Bacteroides_fragilis	0.465	-1
Oceanimonas_sp._GK1	Bacteroides_fragilis	0.454	-1
Oceanobacillus_iheyensis	Bacteroides_fragilis	0.457	-1
Ca.Phytoplasma_asteris	Bacteroides_fragilis	0.438	-1
Opitutus_terrae	Bacteroides_fragilis	0.493	-1
Oscillibacter_valericigenes	Bacteroides_fragilis	0.529	-1
Paenibacillus_mucilaginosus	Bacteroides_fragilis	0.559	-1
Paenibacillus_sp._Y412MC10	Bacteroides_fragilis	0.407	-1
Peptoclostridium_difficile	Bacteroides_fragilis	0.622	-1
Phaeobacter_gallaeciensis	Bacteroides_fragilis	0.462	-1
Phenylobacterium_zucineum	Bacteroides_fragilis	0.463	-1

Table 2.6 continued

Propionibacterium_acidipropionici	Bacteroides_fragilis	0.554	-1
Propionibacterium_avidum	Bacteroides_fragilis	0.551	-1
Propionibacterium_freudenreichii	Bacteroides_fragilis	0.575	-1
Propionibacterium_propionicum	Bacteroides_fragilis	0.562	-1
Pseudomonas_aeruginosa	Bacteroides_fragilis	0.443	-1
Pseudomonas_brassicacearum	Bacteroides_fragilis	0.621	-1
Pseudonocardia_dioxanivorans	Bacteroides_fragilis	0.477	-1
Psychrobacter_arcticus	Bacteroides_fragilis	0.439	-1
Psychrobacter_sp._PRwf-1	Bacteroides_fragilis	0.421	-1
Pyrococcus_sp._ST04	Bacteroides_fragilis	0.424	-1
Rhizobium_leguminosarum	Bacteroides_fragilis	0.548	-1
Rhodobacter_sphaeroides	Bacteroides_fragilis	0.431	-1
Rhodospirillum_photometricum	Bacteroides_fragilis	0.480	-1
Rickettsia_heilongjiangensis	Bacteroides_fragilis	0.445	-1
Roseburia_intestinalis	Bacteroides_fragilis	0.504	-1
Roseiflexus_castenholzii	Bacteroides_fragilis	0.482	-1
Ruminococcus_bromii	Bacteroides_fragilis	0.543	-1
Ruminococcus_obeum	Bacteroides_fragilis	0.403	-1
Ruminococcus_sp._SR1SLASH5	Bacteroides_fragilis	0.556	-1
Ruminococcus_torques	Bacteroides_fragilis	0.438	-1
Sanguibacter_keddieii	Bacteroides_fragilis	0.428	-1
Selenomonas_sputigena	Bacteroides_fragilis	0.493	-1
Shewanella_sediminis	Bacteroides_fragilis	0.526	-1
Shewanella_woodyi	Bacteroides_fragilis	0.554	-1
Sideroxydans_lithotrophicus	Bacteroides_fragilis	0.541	-1
Simkania_negevensis	Bacteroides_fragilis	0.440	-1

Table 2.6 continued

Singulisphaera_acidiphila	Bacteroides_fragilis	0.463	-1
Sinorhizobium_fredii	Bacteroides_fragilis	0.565	-1
Sinorhizobium_meliloti	Bacteroides_fragilis	0.446	-1
Sphingomonas_sp._MM-1	Bacteroides_fragilis	0.467	-1
Sphingomonas_wittichii	Bacteroides_fragilis	0.447	-1
Sphingopyxis_alaskensis	Bacteroides_fragilis	0.491	-1
Spirochaeta_smaragdinae	Bacteroides_fragilis	0.542	-1
Spiroplasma_chrysopicola	Bacteroides_fragilis	0.412	-1
Spiroplasma_taiwanense	Bacteroides_fragilis	0.495	-1
Stackebrandtia_nassauensis	Bacteroides_fragilis	0.421	-1
Staphylococcus_lugdunensis	Bacteroides_fragilis	0.536	-1
Streptococcus_lutetiensis	Bacteroides_fragilis	0.482	-1
Streptomyces_rapamycinicus	Bacteroides_fragilis	0.521	-1
Streptomyces_violaceusniger	Bacteroides_fragilis	0.409	-1
Streptosporangium_roseum	Bacteroides_fragilis	0.596	-1
Sulfolobus_islandicus	Bacteroides_fragilis	0.486	-1
Sulfurimonas_autotrophica	Bacteroides_fragilis	0.645	-1
Synechococcus_sp._PCC_6312	Bacteroides_fragilis	0.506	-1
Synechococcus_sp._PCC_7502	Bacteroides_fragilis	0.450	-1
Terriglobus_roseus	Bacteroides_fragilis	0.449	-1
Thalassolituus_oleivorans	Bacteroides_fragilis	0.409	-1
Thermoanaerobacter_sp._X514	Bacteroides_fragilis	0.616	-1
Thermobifida_fusca	Bacteroides_fragilis	0.425	-1
Thermococcus_sp._AM4	Bacteroides_fragilis	0.447	-1
Thermocrinis_albus	Bacteroides_fragilis	0.548	-1
Thermodesulfobium_narugense	Bacteroides_fragilis	0.414	-1

Table 2.6 continued

Thermodesulfovibrio_yellowstonii	Bacteroides_fragilis	0.497	-1
Thermotoga_naphthophila	Bacteroides_fragilis	0.413	-1
Thermovibrio_ammonificans	Bacteroides_fragilis	0.444	-1
Thioalkalimicrobium_cyclicum	Bacteroides_fragilis	0.408	-1
Tistrella_mobilis	Bacteroides_fragilis	0.618	-1
Treponema_azotonutricium	Bacteroides_fragilis	0.510	-1
Treponema_caldaria	Bacteroides_fragilis	0.440	-1
Treponema_succinifaciens	Bacteroides_fragilis	0.479	-1
Turneriella_parva	Bacteroides_fragilis	0.536	-1
Vibrio_campbellii	Bacteroides_fragilis	0.409	-1
Xanthomonas_campestris	Bacteroides_fragilis	0.563	-1
butyrate-producing_bacterium	Bacteroides_fragilis	0.426	-1

CHAPTER 3
DYNAMIC CHANGES OF HUMAN GUT MICROBIOTA DURING AGING PROGRESSION

3.1 Introduction

The human gut is an eco-system containing more than one hundred trillion microbes, and plays an important role in the host health[130]. The long-term natural selection acting on both the hosts and microbes leads to a relatively stable structure of the intestinal microbiota, which has been proved finally promoting mutual cooperation and functional stability of this inherently complex ecosystem[18]. Factors influencing the variation of gut microbial community in different individuals include environment, diet, host genetics and the pathological conditions of the hosts[17, 80, 35, 131, 79]. Aging is a process capturing many aspects of the biological variation of the human body, which was accompanied by an increased incidence of infection and functional decline in the gut of elderly individuals[45].

Several previous studies have reported age-related changes of human gut microbiota[46, 47, 48, 49, 50, 51, 52, 53, 54]. By culturing microbes, Hopkins et al. found larger number of *Enterobacteria* in children's fecal than in adults[46]. Yatsunenko et al. found the number of *Bifidobacterium* declined as ages of the hosts increased using 16S rRNA sequencing[48]. Odamaki et al. revealed that there was an increasing proportion of *Bacteroides*, *Eubacterium* and *Clostridiaceae* accompanying aging; while *Enterobacteriaceae* were enriched in elderly and infant; *Bifidobacterium* were more abundant in infants; *Lachnospiraceae* were more abundant in adults[47]. Stewart et al. discovered L-lactate dehydrogenase major in milk fermentation declined and transketolase major in the metabolism of fiber increased over the first years of life[50] using whole genome sequencing. In these above mentioned studies, various supervised machine learning algorithms

have been applied to effectively identify taxonomic and functional signatures for aging-related variations of gut microbiota, including multi-group Spearman rank correlation, Random Forest[48], comparative analysis with permutational analysis of variance (PERMANOVA)[46, 47, 49, 54], and frequency-inverse document frequency and minimum-redundancy maximum-relevance[51].

As an exploration, we proposed to apply an unsupervised machine learning algorithm to reveal the existence of aging-related progression of microbial community, and those bacteria genera associated with this progression. Sample Progression Discovery (SPD) as an unsupervised machine learning algorithm was adopted here. SPD was previously developed to reveal the progressively transforming patterns of gene expression, which could be applied to identify those biological progressions of various biological systems and processes[81]. The idea of SPD has been sequentially applied to gene expression analysis of microarray data[81], and then extended to the analysis of flow cytometry data[82] and the analysis of single-cell RNA-seq data[132]. In this part of this dissertation, we applied SPD on community profiles of human gut microbiota samples extracted from 16S rRNA sequencing data. These samples cover various age periods ranging from new-born babies to centenarians. SPD successfully recapitulated the underlying aging progression of the data in an unsupervised fashion, which sorted the gut microbiota samples in an order consistent to the host ages. Additionally, SPD identified those bacteria genera associated with this aging-related progression of the gut microbiota. Overall, these findings proved the existence of an aging progression in human gut microbial community, and identifies some important bacteria genera that could characterize the aging progression of gut microbiota.

3.2 Data and methods

3.2.1 Data and data annotation

There are 371 samples of subjects included in this study, ranging from new-born babies to centenarians, which have been described in a previous publication[47]. We downloaded

the raw data of 16S rRNA data following the accession number DRA004160 from DNA data bank of Japan. Three samples were eliminated because that only one end of paired-end reads could be found. 16S rRNA data processing was performed using Mothur pipeline[133]. Those raw reads with average quality score < 25 or read length < 150bp were classified as low quality and then filtered out. The criteria of minimum length of reads was set as 150bp for the reason that the average overlap region of paired reads was about 150bp. Since the number of reads in each sample was distributed in a Gaussian shape $(8,734 \pm 2,748)$, we could conclude that all the 368 samples were sequenced in a normal depth. Those reads passing quality control and also with both paired ends were merged as sequences, while those low-quality reads failed to pass quality control or with only one end which was supposed to be a pair were discarded. Then the merged sequences were aligned against Silva reference database version 132[134] to annotate the taxonomical composition. The threshold of bootstrap confidence value for the alignment was set 80% (80% identity) during 100 iterations. The taxonomic composition at genus level could be revealed according to the alignment result, which resulted in 368 genera in total.

3.2.2 Feature matrix

The genus abundance matrix was defined as $N = \{n_{ij}\}$, wherein n_{ij} represents the number of reads from sample i classified into genus j. 119 genera with extremely abundance were filtered out, and three genera with unclear annotations were combined into one genus cluster named "unclassified", after which 247 features were obtained for further analysis. To eliminate the influence of various sequencing depth of different samples, we transferred the genus abundance matrix into a relative abundance matrix $F = \{f_{ij}\}$, where $f_{ij} = n_{ij} / \sum_{k=1}^{247} n_{i,k}$. One sample from subject "Japanese 320" with abnormally high proportion of *Pseudomonas* was filtered out. Finally, a 367×247 relative abundance matrix F was obtained for further analysis.

With decent numbers of samples in all age groups, the genus relative abundance of pop-

ulation in each age group was estimated as the mean value of samples in the corresponding group. This step partially reduced the sparsity of the data matrix and the variations of individual samples. Age periods were defined regarding the physiological transition of the hosts, wherein the new-born babies were grouped according to their weaning status and the adults were grouped by decade. The number of samples in each age group was depicted in Table 3.1.

3.3 Results

3.3.1 Data annotation and samples overview

As initially raw data, a total of 3.2 million high-quality 16S rRNA sequences were obtained from 368 samples[47], with $8,734 \pm 2,748$ (mean $\pm$ deviation) reads per sample. The 16S rRNA sequences were binned into 366 genera according to the common-used pipeline of Mothur[133] with SILVA[134] as the reference database (see Data and methods for more details). 119 genera with extremely low abundance were removed. The total number of sequences annotated to these genera only accounted for 0.01% of all the sequences. Also, the sample 'Japanese 320' was excluded for its abnormally high proportion of *Pseudomonas*, which pointed to either pathological disorder of this individual or normal sampling. After all these preprocesses, a relative abundance matrix of the 247 genera across the 367 samples were derived, which was used as the basis for further analyses. For the purpose of revealing the existence of age-related progression of gut microbiota, all the samples were divided into 14 age groups by considering the physical transformation of different body periods. Accordingly, adults were grouped by decade, while new-born babies were grouped based on their weaning status (Table 3.1). Except the group of centenarians, there were at least 10 samples in each age group.

Principle component analysis (PCA) was performed here to visualize the taxonomic patterns of these samples by transforming the high dimensional raw data into a three-dimension space. PCA was applied on the relative abundance matrix of the 247 genera

Table 3.1: Samples were grouped into 14 age-segment groups. The first three groups of new-born babies were classified regarding their weaning status, i.e. before weaning, weaning and after weaning separately. Other samples were grouped by decade.

Group	Age segmentation	Number of samples	Female	Male
1	(0, 0.4]	10	6	4
2	(0.4, 1.2]	12	4	8
3	(1.2, 3]	19	9	10
4	(3, 9]	14	8	6
5	(9, 19]	10	3	7
6	(19, 29]	40	24	16
7	(29, 39]	88	43	45
8	(39, 49]	34	21	13
9	(49, 59]	25	13	12
10	(59, 69]	28	17	11
11	(69, 79]	15	10	5
12	(79, 89]	48	32	16
13	(89, 99]	19	15	4
14	≥ 100	5	5	0

across the 367 samples. The top three principle components with the highest explanation power explained 33.17%, 15.09% and 10.32% of the original data variance, respectively. As Figure 3.1 shows, those samples from children younger than three years old scattered loosely, which means they are quite different from each other. This observation confirmed the finding of the previous literature[48], which discovered that interpersonal variation decreased as the hosts become older. Nevertheless, the samples did not gather into distinct groups when analyzed using this linear analysis method.

3.3.2 Age-related variation of gut microbiota revealed by supervised methods

Two traditional supervised statistical approaches was applied in a univariate fashion to identify the age-related variation of the gut microbiota. First, permutational one-way ANOVA test[135] was applied on the genus relative abundance matrix to find those genera with abundances significantly varied across different age groups. The abundances of forty three genera showed significant difference in different age groups with $P < 0.001$ during 1000

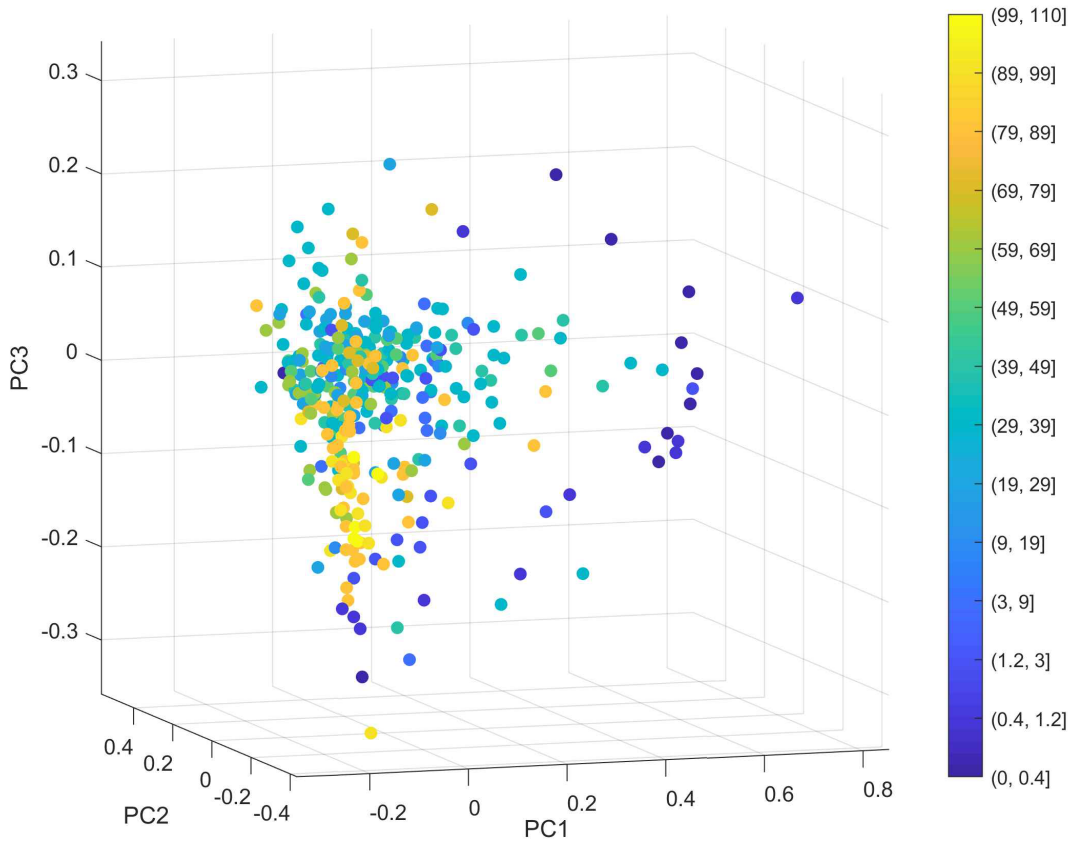

Figure 3.1: Sample overview using PCA. Using the relative abundance of 247 genera across all the 367 samples as input, we linearly transformed and visualized the data in a three-dimensional space. Each sample is represented by one dot, colored according to age. Samples from children younger than three (the dark blue dots) scattered most distantly, while older age groups were mixed together in the PCA space.

times of randomization. Herein the P values have been adjusted using Bonferroni correction (see more details in Table 3.2). Spearman correlation was then applied to identify those genera co-varying with aging. 17 genera were identified positively correlating with aging and one genus negatively correlating with aging (Table 3.3). These findings agreed with multiple previous literatures, which also revealed the variation of individual genus in the gut microbial community during the host aging[46, 47, 49, 54]. Another further question naturally arose as to whether the gut microbial community shift continuously during aging as a whole.

Table 3.2: Significant genera from Permutational one-way ANOVA analysis

Alistipes
Anaerofilum
Anaerostipes
Bifidobacterium
Bilophila
Blautia
Butyricicoccus
Butyricimonas
Christensenellaceae_R-7_group
Christensenellaceae_ge
Cloacibacillus
Clostridium_sensu_stricto_1
Coprococcus_1
Corynebacterium
Desulfovibrio
Dorea

Table 3.2 continued

Escherichia-Shigella
Family_XIII_UCG-001
Fusicatenibacter
GCA-900066225
Intestinibacter
Intestinimonas
Izimaplasmatales_ge
Lachnospira
Lachnospiraceae_ge
Methanobrevibacter
Negativibacillus
Odoribacter
Parabacteroides
Phascolarctobacterium
Rhodanobacter
Ruminiclostridium_5
Ruminiclostridium_9
Ruminococcaceae_NK4A214_group
Ruminococcaceae_UCG-002
Ruminococcaceae_UCG-005
Ruminococcaceae_UCG-010
Ruminococcaceae_ge
Ruminococcus_2
Sellimonas
Subdoligranulum
Sutterella

Table 3.2 continued

Veillonella

Table 3.3: Genera correlated with aging with Spearman correlation

Postive	Negative
Alistipes	Bifidobacterium
Butyricimonas	
Christensenellaceae_R-7_group	
Christensenellaceae_ge	
Cloacibacillus	
Desulfovibrio	
GCA-900066225	
Intestinimonas	
Odoribacter	
Parabacteroides	
Phascolarctobacterium	
Ruminiclostridium_9	
Ruminococcaceae_NK4A214_group	
Ruminococcaceae_UCG-002	
Ruminococcaceae_UCG-005	
Ruminococcaceae_UCG-010	
Ruminococcaceae_ge	

3.3.3 Aging progression of gut microbiota revealed by unsupervised analysis

An unsupervised method SPD was applied to analyze the gut microbiota data in a multivariate fashion, which is totally different from the previous supervised univariate methods that search for features co-varying with aging. The averages of genus relative abundance of samples in each age group, which is a 247×14 matrix, was input to SPD. The scale effect was eliminated by normalizing the relative abundance of each feature across samples. A minimum spanning tree (MST) was constructed for each of the genus features based on Euclidean distance. This MST represented a putative progression ordering among the 14

sample groups. The 247 resulting MSTs for those 247 genera were cross compared with each other to examine whether there is a relatively dominant progression ordering fitted well by multiple genera among the samples. SPD used a progression similarity matrix to summarize the results of these comparisons, wherein each element of the matrix counted the number of progression orderings that the two corresponding genera both fit well with. The result was shown in Figure 3.2a and part of the highlighted area of this matrix was magnified in Figure 3.2b. It could be observed that a subset of 35 genera (Table 3.4) fitting well with a common set of putative progression orderings. An overall minimal spanning tree could be constructed using this subset of 35 genera, which represent a common progression ordering (shown in Figure 3.2c). One age was represented by one node of the tree in Figure 3.2c. To assist the visualization, nodes were labeled and colored according to the order of their real age groups. Noticeable, for determining the structure of the tree, the age information was not used.SPD aimed to identify a progression ordering among the samples, represented by an overall minimal spanning tree, and discover those features exhibiting gradual changes respecting this progression. The age progression ordering across the 14 sample groups was recapitulated by the overall minimal spanning in Figure 3.2c.

Table 3.4: Critical genera identified by SPD

Critical genera
Allisonella
Acidocella
Oscillospira
Angelakisella
Parvimonas
Ruminiclostridium_9
Ruminococcaceae_UCG-003
Anaerotruncus

Table 3.4 continued

Clostridiales_vadinBB60_group_ge
Cloacibacillus
Corynebacterium
GCA-900066225
Desulfovibrio
Christensenellaceae_R-7_group
Harryflintia
Ruminococcaceae_NK4A214_group
DTU014_ge
Rikenellaceae_RC9_gut_group
Bilophila
GCA-900066755
Lactobacillus
Tyzzerella
Butyricimonas
Odoribacter
Phascolarctobacterium
Butyrivibrio
Prevotellaceae_UCG-003
Senegalimassilia
Oxalobacter
Peptococcus
Parascardovia
Rhodanobacter
Cellulosilyticum
Epulopiscium

Table 3.4 continued

Pyramidobacter

These sample groups could be further classified into four larger groups, i.e. Centenarians, Elderly, Adults, and Children and teenagers. The order of these larger sample groups on this minimal spanning tree perfectly matched with the aging order of these sample groups. This finding is interesting since it discovered that there existed an aging progression of the human gut microbiota, based on the truth that SPD recovered the correct ordering of aging progression using the genus relative abundance without any prior knowledge.

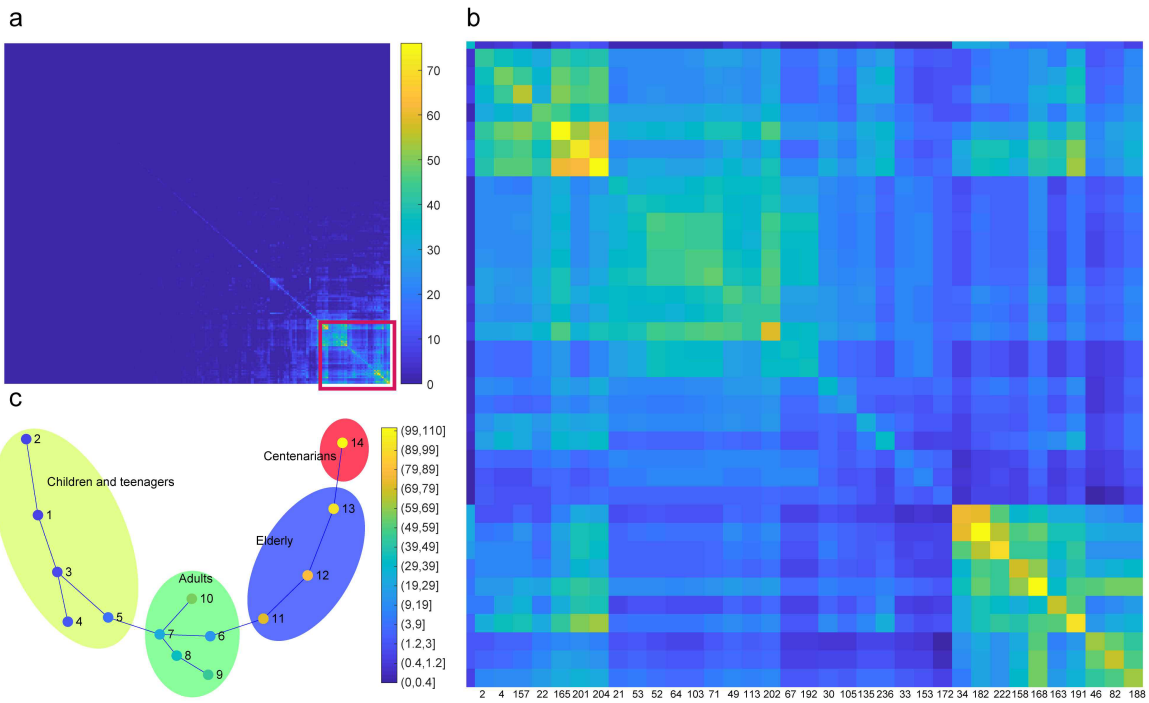

Figure 3.2: SPD recovered aging progression with taxonomical composition of human gut microbiota. (**a**) Progression similarity matrix for all genera, with each element counting the number of progression orderings the two corresponding genera shared. (**b**) We manually picked the highlighted area from (**a**). These selected genera were consistent with a common set of putative progression orderings. (**c**) An overall minimal spanning tree of the 14 age groups based on the selected genera. Each node represents one age group.

3.3.4 35 critical genera underlying the aging progression of gut microbiota

We further examined those 35 identified genera that contributed to the aging progression of gut microbiota. Compared to previous methods mentioned above, out of the 35 genera, 11 were detected as significant features in the permutational one-way ANOVA analysis with adjusted $Pvalue < 0.001$. None of them was detected as correlated with aging by Spearman rank correlation analysis. For remaining 24 genera only detected by SPD, a few have been previously reported in other literature, such as *Butyrivibrio*, *Oxalobacter*, *Lactobacillus* which have been experimentally demonstrated associating with aging[136, 137, 138], as well as *Prevotellaceae* which has been reported with lower presence in the gut microbiota of centenarians[139].

Regarding the varying trend of these critical genera, among the 35 genera selected by SPD according to the progression similarity, only 9 of them monotonically varies along with aging, while the rest of them increased at the beginning while then decreased in different age periods (Figure 3.3). This illustrated one advantage of SPD, which was designed to find those features exhibiting gradual changes respecting a common underlying progression pattern, and the gradual changes were not limited to be monotonic and also include those not monotonic changes. Therefore, this analysis was able to discover all those genera gradually changing without abrupt fluctuations during aging. Extensive literature review of these 35 genera has been performed, and a lot of previous reports of the functional relevance of these genera has been published.

Figure3.4 shows those genera with abundances increased with respect to aging, but decreased in the extremely elderly subjects. Within these genera, *Lactobacillus* species are commonly used probiotics[140]. Species in genus *Oscillospira* are central to the human gut microbiota for degrading fibers[141], and have been frequently reported enriched in lean subjects compared to those obese subjects [142, 143, 144, 145]. Genus *Oxalobacter* is responsible for degrading oxalate in the gut, and it has been experimentally demonstrated existing in the gut of almost all young individuals, but these bacterium may later get

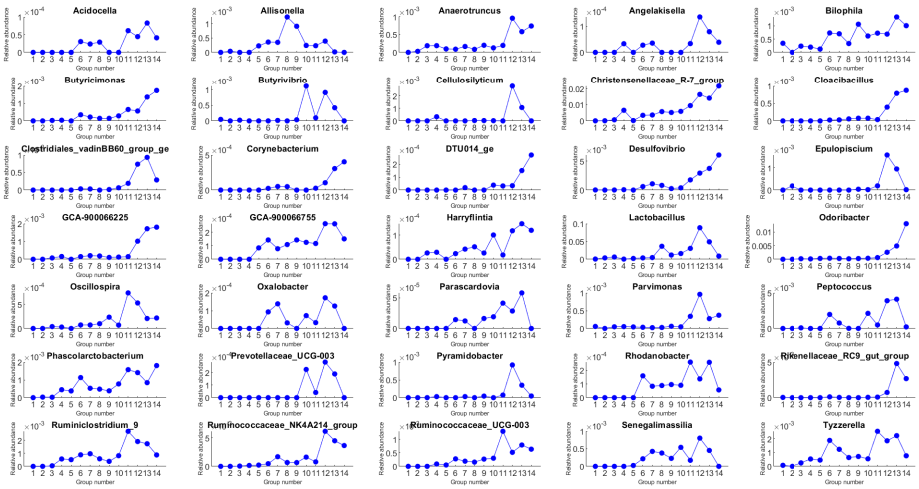

Figure 3.3: The relative abundance of all the 35 critical genera across different age groups.

lost during aging[136]. *Prevotellaceae* is commonly found in the gastric system of people who maintain a diet high in carbohydrates and low in animal fats[146] and could be lost in centenarians[139]. Researchers also found that there was an increased abundance of genus *Prevotellaceae* in the guts of healthy individuals than those people with Parkinsons disease[147]. *Parascardovia* is a genus within the family *Bifidobacteriaceae*. This family has been shown benefiting the host health in multiple ways[148]. Species within genus *Butyrivibrio* have been proved experimentally as butyrate producing bacteria, while butyrate is a preferred energy source for colonic epithelial cells and has been demonstrated to play an important role in maintaining colonic health of hosts[149]. Integrating all these findings, it could be concluded that these genera are health beneficial ones. The decrease of these beneficial genera in the elderly age groups, especially centenarians, maybe manifestation of or causal associations to decline of health in those age groups.

In contrast, genera in Figure 3.5 showed generally monotonically increasing patterns respecting the whole aging progression. Literature review of these genera led us to conclude that these genera are most likely to be health harmful. Genus *Parvimonas* was enriched in colorectal cancer compared to the healthy controls[150, 151, 152, 70, 153]. Genus

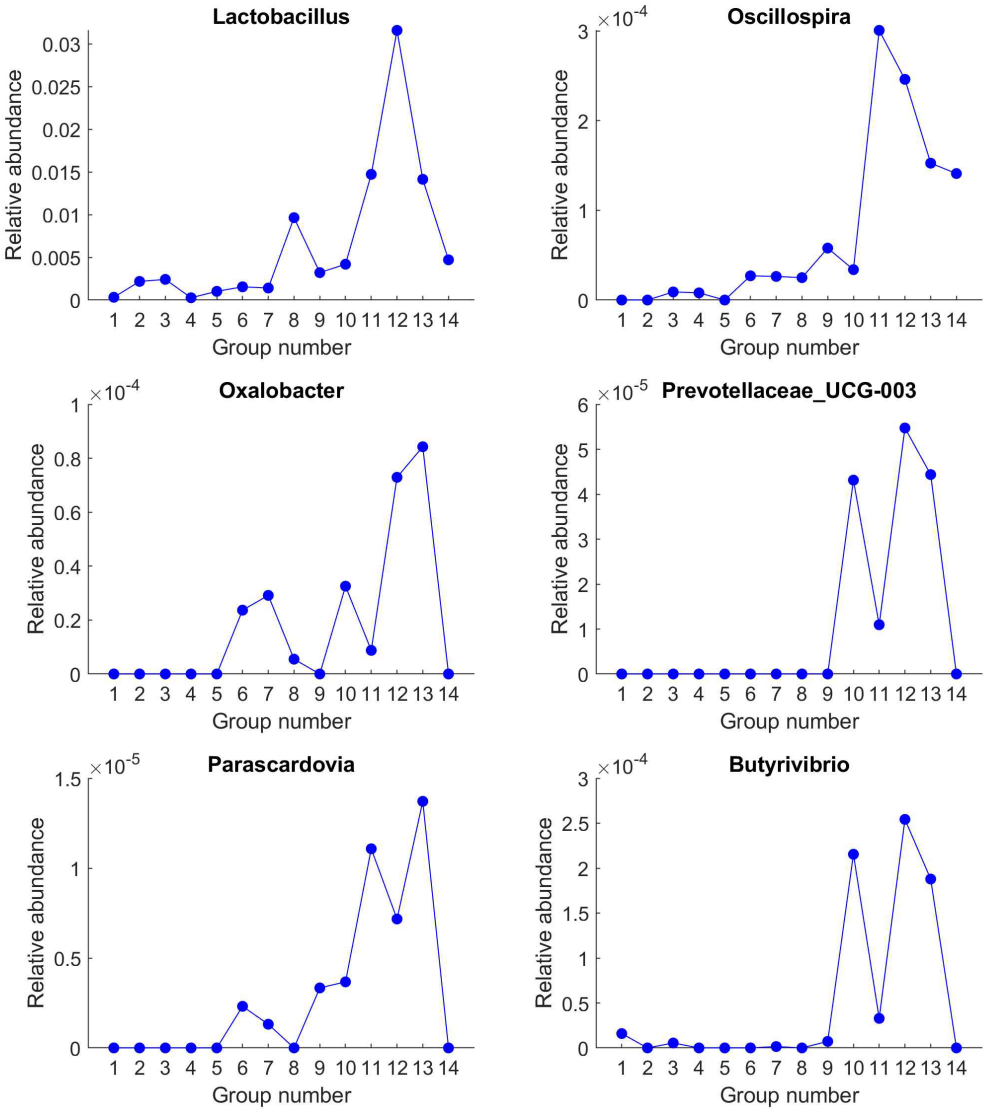

Figure 3.4: Genera that first increased and then decreased during aging, especially sharply decreased the 13th or 14th age groups, or both.

Anaerotruncus has been reported as relatively enriched in patients with age-related macular degeneration[154]. Genus *Corynebacterium* was reported more abundant in the gut of autistic individuals with autism spectrum disorders[155]. Many species within genus *Corynebacterium* have been reported involved in human and animal diseases[156]. Genus GCA-900066225 is one representative genus in the family *Lachnospiraceae*, which has been reported to be associated with the stress of the host, ulcerative colitis, as well as Crohn's and celiac disease[157]. Genus *Desulfovibrio* produce hydrogen sulfide using sulfate as the electron acceptor, and these sulfate-reducing bacteria are often reported associating with the host inflammation[158, 159]. Strain *Bilophila wadsworthia* within genus *Bilophila* has been reported as causing systemic inflammation in specific-pathogen-free mice[160]. Species within genus *Odoribacter* has been reported enriched in tumor-bearing mice[66]. Genus *Butyricimonas* was more abundant in those subjects suffering from systolic blood pressure, high rectal temperature, and with a significantly lower physical activity score[161]. Overall, these monotonically increasing genera were often reported to relate with host diseases.

All these prior literature of these identified genera led to one interesting finding. Many of the genera first increasing and then decreasing were previously demonstrated as health beneficial, whereas most of the monotonically increasing genera along with aging were frequently identified as disease related genera. Specially, when individuals turn elderly beyond 90s, their guts tend to lose some beneficial genera while get some potentially harmful genera.

3.4 Discussion

Since the variation of gut microbiota is intensively related to the health status of the hosts, in order to exclude other influences, an ideal dataset for examining aging related changes of human gut microbiota should be collected from healthy subjects form different age groups. Unfortunately, we do not know the healthy status of individuals in this study, because

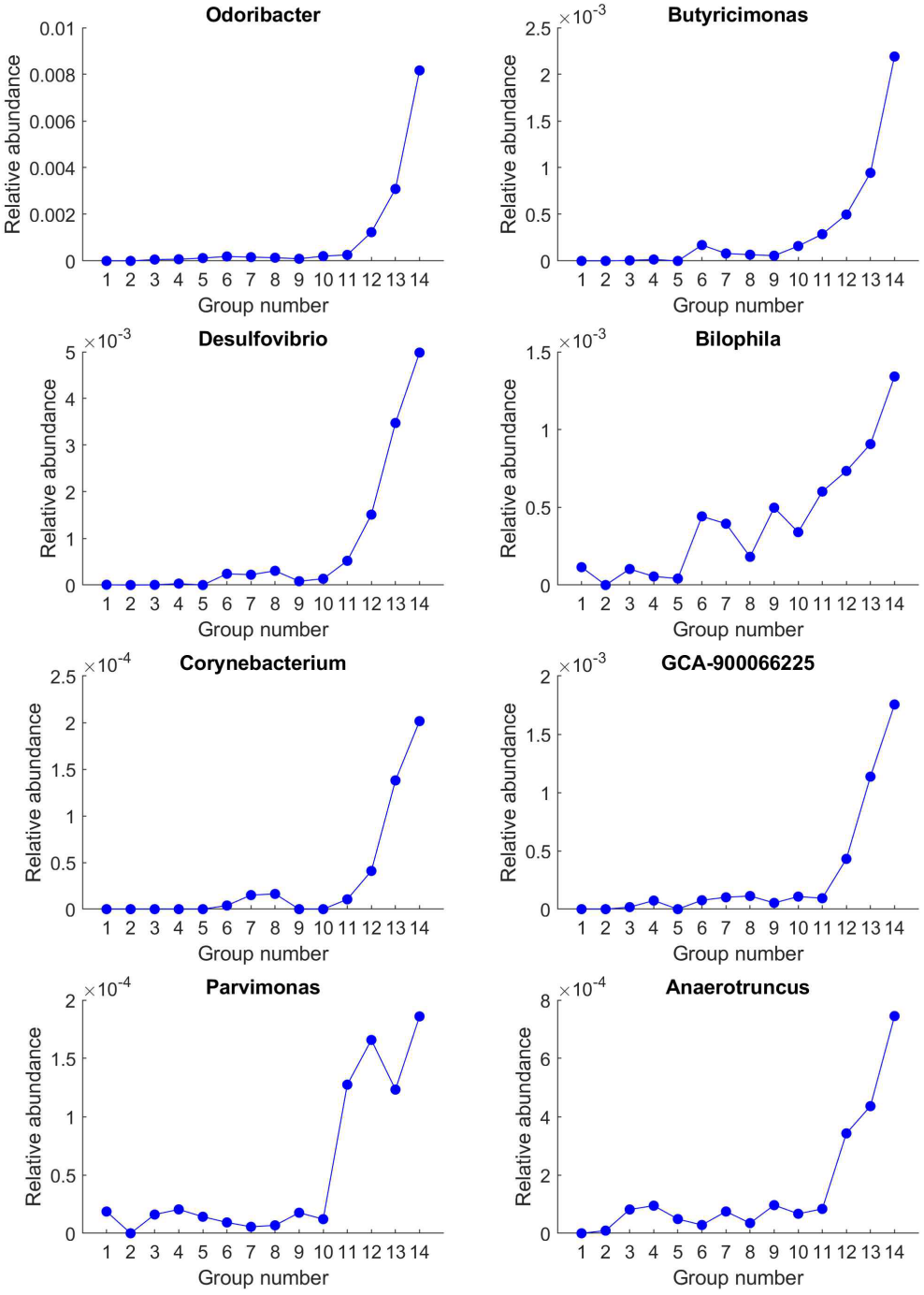

Figure 3.5: Genera that exhibited general increasing patterns during aging.

that the data we used here were obtained from a previously published paper[47] which did not include the detail of health information for these samples. During our literature search on age-related variation of the human gut microbiota, we realised that the health status of individuals in such studies are often not reported in multiple previously published papers[47, 48, 162, 51, 163].

We performed extra analysis to gauge the health status of those samples used in this part. We retrieved multiple previous datasets on the human gut microbiome of hosts suffering from different diseases[80, 79, 32, 78, 33, 43], which have been detailed in Chapter 2 of this dissertation. We obtained the relative abundance of the human gut microbial genera of each sample in these previous datasets. The abundance distribution of these genera could be visualized for both disease samples and healthy controls. It could be clearly observed that some genera were more abundant in the disease samples compared to those healthy controls, and the majority of these genera have been demonstrated as opportunistic pathogens of the human gut[164, 165, 166, 167, 168, 169, 170, 171, 172, 173]. According to Figure 3.6, those disease-enriched genera typically showed higher variance and higher abundance in disease samples than in those healthy ones (first and second columns of Figure 3.6), while all of these disease-enriched genera exhibit very low abundance in the dataset used in this part of analysis (third column of Figure 3.6). This observation implied that the samples in the current dataset are dissimilar to the diseases samples while more similar to the healthy samples in the previous datasets. Thus this comparison demonstrated that the most of samples in this part of analysis were derived from healthy individuals.

Operational taxonomic unit (OTU) is another commonly used classification unit for 16S rRNA sequencing data analysis, which allows for classifying 16S rRNA sequences into features at a finer resolution compared to the previously used genus level classification. we applied the progression analysis to the OTU level features, the result of which confirmed our observations in the genus level analysis. In detail, we clustered sequences with similarity threshold 0.97, which resulted in 4,663 OTUs. Those OTUs with extremely

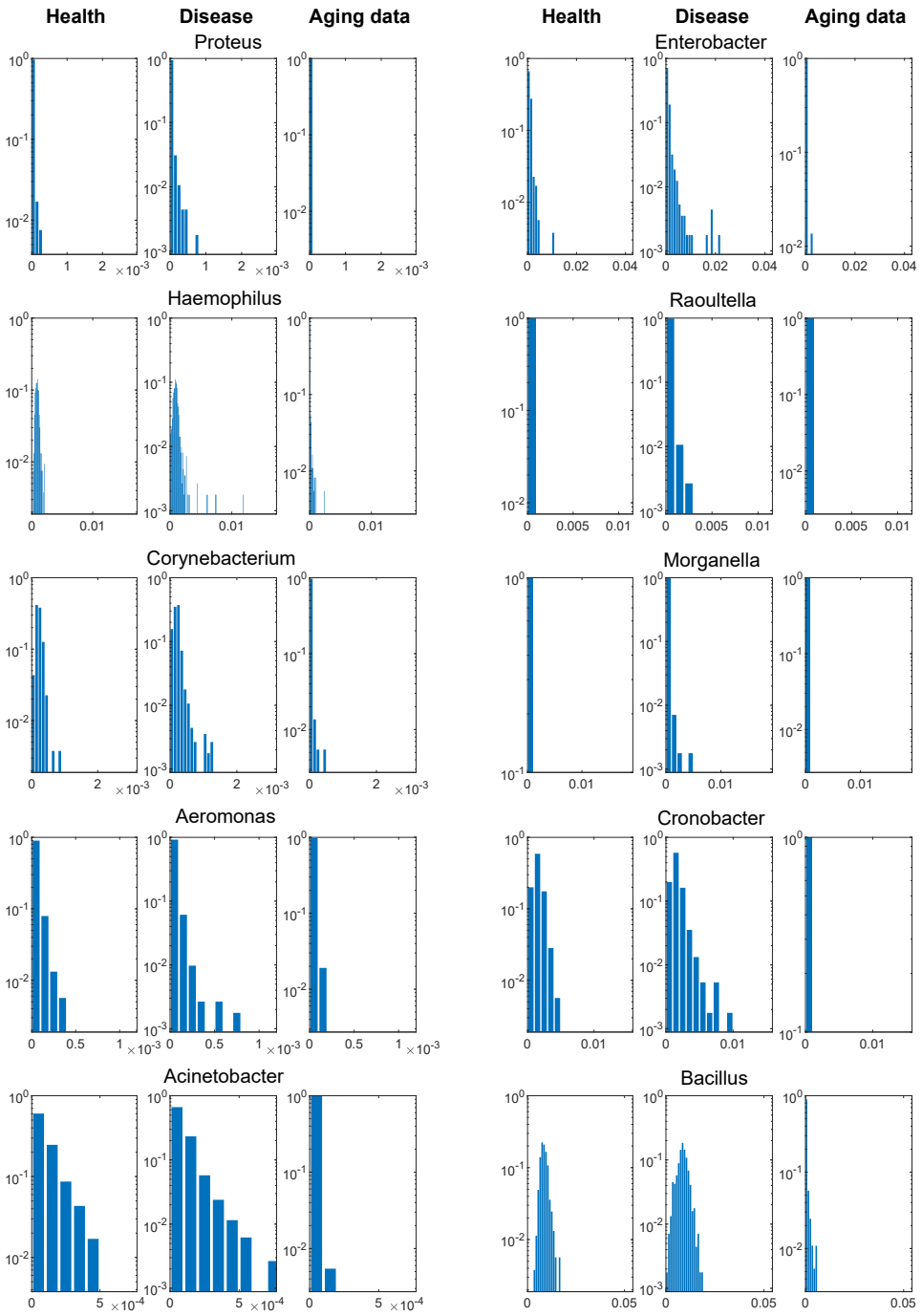

Figure 3.6: Frequency histogram of the relative abundance of disease-enriched genera distributed in different kinds of samples in our previous studies and the current dataset. All the value bins along the x-axis are consistent and all the relative abundance values were log transformed before being binned. We could see that the distribution of the samples we included in this paper is more similar to the healthy samples and exhibit lower abundance compared to the disease samples.

low abundances were filtered, and there were 1,229 OTUs passing this filtering step. The input for SPD analysis was the averages of the relative abundances of the remaining OTUs for each age group. The result showed that the progression analysis based on OTU features could partially recapitulate the correct order of the age groups (Figure 3.7), but not as good as the result from genus level analysis as shown in Figure3.2c. Nevertheless, it reassured the existence of an aging progression of human gut microbiota, which could be demonstrated by the analyses at both OTU level and genus level.

For quantifying species diversity in the metagenomics literature, the alpha diversity and the beta diversity are very popular metrics. Herein, we calculated the alpha diversity and the beta diversity of gut microbiota in each age group based on the averages of genus relative abundance of samples in the corresponding group. We chose Shannon index to quantify the alpha diversity and Bray-Curtis dissimilarity between different age groups for quantifying the beta diversity. As shown in Figure 3.8, the alpha diversity of each individual age group steadily increases as a function of aging, while experienced a steep drop in the extremely elderly age group [99, 110]. This observation agreed with the results shown in Figure3.4, where multiple aging-related genera significant decreased in the extremely elderly age group. Different from alpha diversity, the beta diversity was usually applied to quantify the dissimilarity between different age groups (Figure 3.9). When focusing on the variation of beta diversities between neighboring age groups, it could be observed that the beta diversity between groups [2, 3] and between groups [13, 14] were notably larger than the beta diversities between other neighboring age groups. The specially larger distinction between group 2 (weaning) and group 3 (weaned) could be explained by the transformation of weaning status, which is often accompanied by drastic dietary changes. However, those samples in group 13 and group 14 are all elderly individuals with continuous ages, thus the alteration of dietary habits can not explain the large dissimilarity between groups 13 and 14 anymore. Therefore, we conjecture that the large dissimilarity between groups 13 and 14 is due to the aging of gut microbiota, manifested in the sudden decrease of multiple

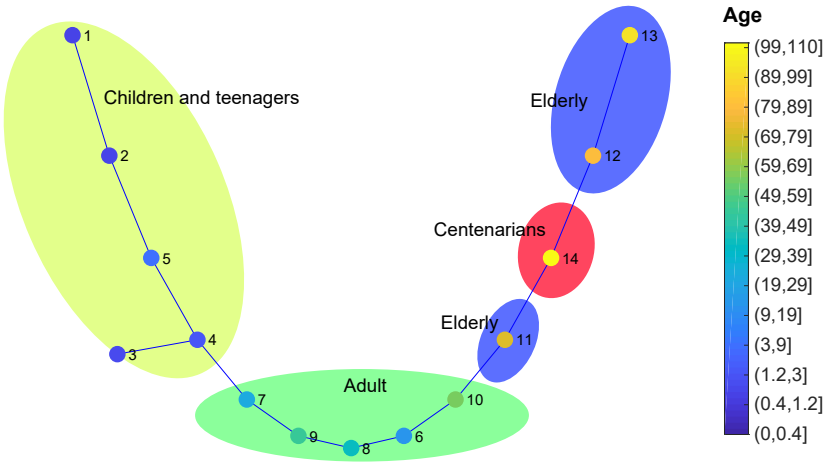

Figure 3.7: The minimal spanning tree generated from SPD based on OTUs.

genera in the extremely elderly samples. Overall, our observations in the analyses of both alpha and beta diversity confirmed our previous observation that the abundance of multiple genera suddenly decreased in those extremely elderly age samples (Figure3.4).

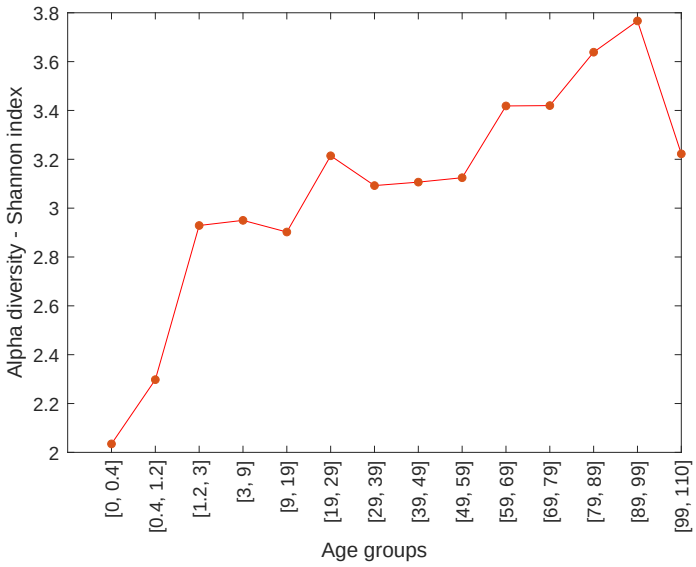

Figure 3.8: The alpha diversity of samples in different age groups. Herein, the alpha diversity is quantified by Shannon index. We could see that the alpha diversity truly decreased for the extremely elderly age groups.

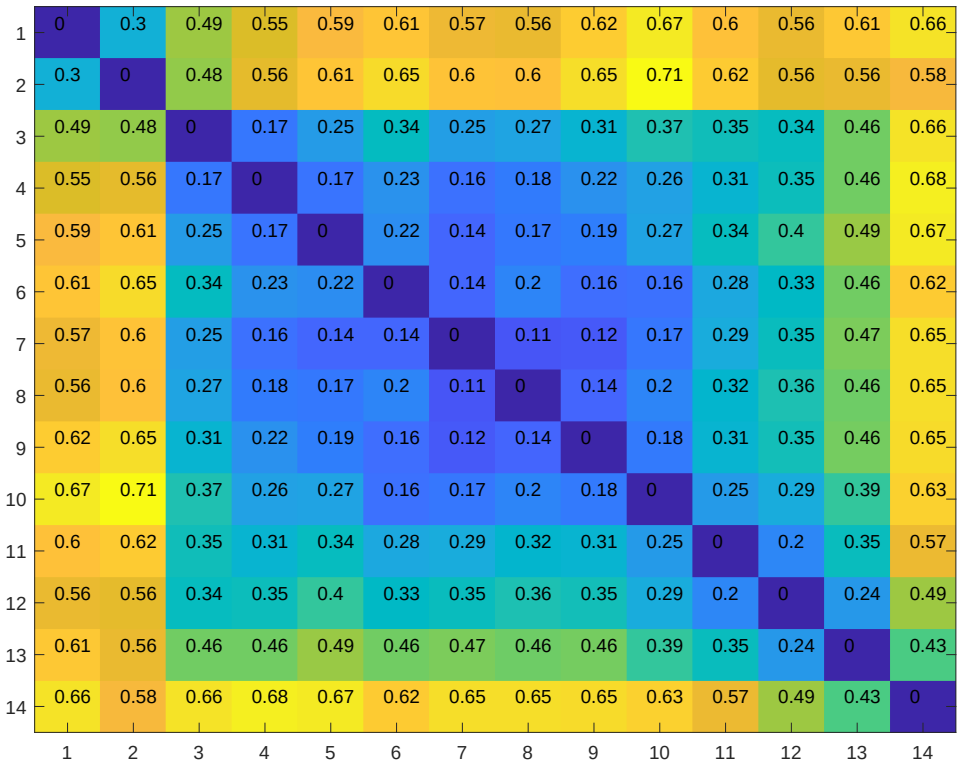

Figure 3.9: The beta diversity between different age groups, which is quantified by Bray-Curtis dissimilarity. The values adjacent to the diagonal line elucidate the dissimilarity between neighboring age groups.

CHAPTER 4
A MACHINE LEARNING TOOL FOR DIAGNOSING DISEASES BASED ON HUMAN GUT MICROBIOTA

4.1 Introduction

IBD is a group of inflammatory conditions of the colon and small intestine that affects over 2.5 million Europeans[84] and 3.1 million Americans[174], and has a notably increasing prevalence in the Asia-Pacific region[175]. An early accurate diagnosis can help clinicians to improve treatment. However, there is no gold standard diagnosis for monitoring quiescent disease in patients with IBD. Moreover, the two major types of IBD, UC and CD[176], have different mechanisms of tissue damage[177] necessitating different treatment strategies. It is clinically critical but usually difficult to identify the specific types of IBD, because there are no biomarkers or clinical tests capable of discriminating CD from UC patients in practice[178]. Even colonoscopy may miss inflammation in some parts of the gastrointestinal tract[179].

The human gut microbiota has been viewed as a relatively forgotten organ, however has been increasingly concerned with an important role in health[18]. Recently the next-generation sequencing (NGS)-based profiling studies of the intestinal microbiome have reinforced the view that the pathogenesis of IBD is closely associated with the unbalanced composition of the microbial community[180, 181, 182]. In contrast to serum biomarkers, fecal biomarkers respond more directly to the changes of the intestinal conditions. With the development of NGS technology and advances in hospital bioinformatics analysis, it is time to propose a diagnostic procedure to discriminate UC and CD from non-IBD colitis, especially based on the current high-throughput NGS data of the human gut microbiome.

In this work, we present a tool, named LightCUD, for discriminating ulcerative colitis

(UC) and Crohn's disease (CD) from non-IBD colitis using the human gut microbiome. LightCUD embodies four high-performance modules, namely, WGS-based health vs IBD module, WGS-based UC vs CD module, 16S-based health vs IBD module and 16S-based UC vs CD module. Each module is composed of a machine learning model and a customized reference database. In details, we used the high-throughput whole-genome sequencing (WGS) data to analyze the microbial composition of gut microbiota samples. These samples were from patients with UC and CD, and healthy controls. The taxonomic profiles of these samples were obtained as feature abundance matrices (FAMs) at strain level for two WGS-based modules and at genus level for two 16S-based modules respectively. We designed a feature selection strategy for all the modules. Also, we compared the performances of five different machine learning algorithms, i.e., logistic regression, random forest, gradient boosting classifier, support vector machine and LightGBM for training each model of corresponding module[94, 93, 92, 91]. The LightGBM-based models performed best. As a result, we established four high-performance lightGBM-based modules, namely, WGS-based health vs IBD module, WGS-based UC vs CD module, 16S-based health vs IBS module and 16S-based UC vs CD module. For the two WGS-based modules, we further optimized the feature/strain sets to improve the modules performance. The result illustrated that 49 strains for WGS-based health vs IBD module and 12 strains for WGS-based UC vs CD module could achieve the best performances. Finally, we constructed and released the tool LightCUD. With 16S rRNA sequencing or WGS data from individual gut microbiota samples as input data, LightCUD predicts the probability of having IBD, and the sample identified as IBD will then be classified as UC or CD.

4.2 Data and methods

As shown in Figure 4.1, we first conducted metagenomics analysis for WGS data of human gut microbiota samples from two types of IBD and healthy controls. Based on the alignment result from last step, we constructed feature profiles at strain level for WGS-

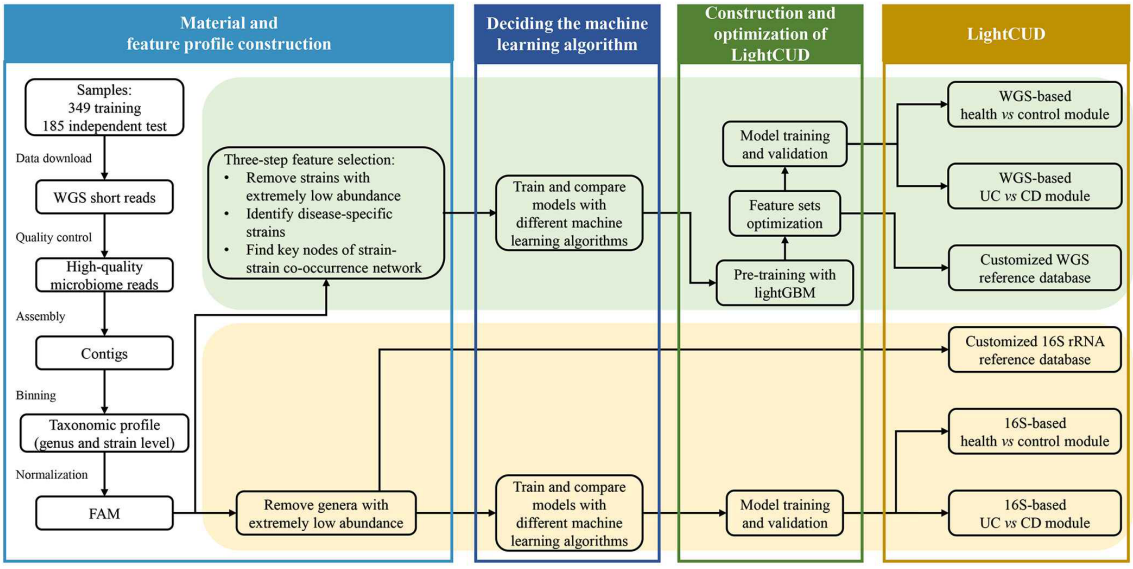

Figure 4.1: The pipeline of data processing and the LightCUD program construction. With WGS raw data of 349 samples, we eliminated the low-quality reads and assembled the remaining reads into contigs. Contigs ¿ 1,000 bp were taxonomically binned into strains and genera. 16S rRNA-based discrimination modules were constructed with genus-level profiles and WGS-based discrimination modules were constructed with strain-level profiles. For the four modules, we designed different feature selection procedures and compared different machine learning algorithms. LightGBM was selected as the core algorithm for modules construction for its best performance. For WGS-based modules, we further optimized the model by shrinking the feature set through pre-training. Finally, a high-performance dual-usage discrimination program LightCUD was successfully constructed. The corresponding reference databases were released along with the prediction modules.

based modules, and at genus level for 16S-based modules. After the well-designed feature selection steps, we selected the LightGBM models to construct the discriminant modules, outperformed the other four machine learning algorithms. We then describe in details about the methods.

4.2.1 Data description

We downloaded a deeply sequenced microbiota data set of 396 human stool samples from public database, which has been described in a previous study of human intestinal tract metagenome[78]. Among the samples, 47 ones labeled with relative health were excluded

because of their uncertainty of being IBD patients or healthy. So totally 349 samples were included in this study, consisting of 201 samples from healthy controls, 127 samples from UC, and 21 samples from CD. 4.68 TB WGS paired-end short reads of these samples were downloaded from NCBI GenBank[95]. We first assembled the original short reads into contigs using InteMAP[183], which was designed as an integrated assembly pipeline for NGS metagenomic short reads. To ensure the validity of further analysis, short reads with low quality and contigs shorter than 1,000 bp were filtered out.

To verify the generalization ability of LightCUD, we also conducted blind validation with an independent data set, which has been described in another study[79]. This data set includes 244 GB short reads of 185 samples with moderate size, including 16 healthy controls and 169 with CD ($> 696MB$ and $< 2000MB$ per sample, filtering out samples in the bottom or top quartiles were filtered out).

4.2.2 Constructing feature profiles for WGS-based and 16S-based modules

NGS techniques enable us to systematically characterize the composition of complex microbial community, such as human gut microbiota. 16S rRNA genes for bacteria were the most commonly used target genes for molecular analysis, which provides fairly consistent taxonomic assignment for a relatively wide range of genera[184]. Although expensive, WGS can theoretically classify taxonomic composition at strain level. WGS-based strain typing is widely used in the epidemiologic analysis of bacterial pathogens in public health, so we developed our program with both WGS-based and 16S rRNA-based modules. In this subsection, we then describe how to construct the strain profiles and genus profiles as FAMs for the WGS-based modules and 16S-based modules separately.

With high-quality contigs, we were then able to recognize the members of the microbial community. To perform taxonomic binning at strain level, we used the 2,712 strain genomes references in NCBI RefSeq[95]. The PhymmBL tool[185] was applied to taxonomic binning, which combined the sequence composition-based method and sequence

alignment algorithm. With Bowtie 2-2.1.0 alignment[110], phylogenetic profiles for a sample were then calculated by counting the number of short reads aligned to each contig. To obtain comparable relative abundances of strains, a correction process for sequencing depth was applied during which the numbers of aligned reads were normalized by the contig length and the number of matches per sample. The resulting values of relative abundances were between 0 and 1. The strain-level taxonomic profiles served as FAMs for WGS-based modules.

In the current study, we also consider such a case that only 16S rRNA data are sequenced for human gut microbiota samples. For this case, we designed 16S-based modules trained with the WGS data herein, and calculated the genus-level taxonomic profiles as FAMs. The relative abundance of each genus in the 16S-based FAM was calculated by adding up the relative abundances of strains belonging to this genus.

For WGS-based modules, we finally annotated 2,661 strains as features with the data set of 349 samples. As we know, a major drawback of WGS analysis is that it is very expensive, mainly because of the large size of whole genome reference database. To address this issue, we optimized the feature set by selecting the most discriminative, so that we could construct a relatively small reference database meanwhile avoiding model overfitting. The features selection process consisted of three steps as follow: 1. Only strains with relative abundances more than 10-6 in at least one sample were reserved. This step was designed to filter out some feature that might be noise information for model training. 2. Group versus group comparative analysis was carried out and strains that significantly passed the Wilcoxon rank sum test ($P < 0.01$) were retained[111]. This analysis is commonly used in metagenomics analysis to identify potentially disease-related taxa. We added this step to select case sensitive strains, and avoid the noise created by insensitive features. 3. The hub nodes/strains in the strain-strain co-occurrence network graphs were selected in an iterative procedure. In co-occurrence networks, nodes were strains and links represented validated strain-strain correlations ($P < 0.05$). The correlations were calculated using the

Spearman correlation in SparCC[112], based on the relative abundances of strains across samples. The most intensively connected strain was picked out as the representative strain (or feature) in each iteration. The selected strain and its strong connected strains ($|R| > 0.4$) were then removed from the graph and the remaining strains were iterated into the next loop of feature selection. This process was kept running until less than three nodes were left in the network graph. This step was designed to select the most representative strains and exclude the strains intensively correlated with the representative strains.

Details of the feature selection process and the detailed parameters are available in the attached R code. We conducted this feature selection process separately for the WGS-based both healthy vs IBD and UC vs CD cases. Finally, we have 320 strains for health vs IBD case, and 159 strains for UC vs CD case, of which 29 overlapped strains were good discriminators for both cases.

For 16S-based module, 508 genera were annotated as features. Since 16S sequencing data is in small size and the alignment is fast, we only need to conduct basic feature selection to exclude noise in feature abundances. In order to eliminate the randomly mapped genera, we required that the relative abundances of selected features are no less than 10-6 in more than 90% samples. Therefore, we have 503 features left for the two 16S-based modules.

4.2.3 Deciding the machine learning algorithm for building LightCUD

With the above optimized feature profiles (strains for WGS-based modules and genera for 16S-based modules), we were able to train the machine learning models of WGS-based modules and 16S-based modules as discrimination methods with five common-used machine learning algorithms and evaluate their performances. The five algorithms are logistic regression, random forest, gradient boosting classifier, support vector machine and LightGBM[94, 93, 92]. Herein the performances were evaluated with five-fold cross validation using the average AUC (area under receiver operating characteristic curve) and AP (av-

erage precision, area under precision-recall curve). AP was adopted as a supplementary measure since the training datasets were unbalanced. For the health vs IBD case, we have 349 samples consisting of 148 IBD and 201 healthy controls. For the UC vs CD case, we have 148 samples containing 127 UC and 21 CD. As shown in Table 4.1, the models built with LightGBM performed overall better than the other four algorithms, with the highest AUC in all the discrimination tasks and the highest AP in three out of four discrimination tasks (WGS-based health vs IBD and UC vs CD, 16S-based health vs IBD and UC vs CD).

Table 4.1: Comparison of model performances built with five different machine learning algorithms. LightGBM performed better than the other four algorithms, with the highest AUC in all the four discrimination tasks (health vs IBD and UC vs CD) and the highest AP in three out of four tasks.

Discrimination tasks	WGS Health vs IBD		WGS UC vs CD		16S Health vs IBD		16S UC vs CD	
Machine learning algorithm	AUC	AP	AUC	AP	AUC	AP	AUC	AP
Logistic regression	0.861	0.781	0.735	0.438	0.857	0.761	0.815	0.559
Random forest	0.899	**0.840**	0.781	0.539	0.873	0.807	0.725	0.520
Gradient boosting classifier	0.796	0.685	0.732	0.331	0.803	0.685	0.858	0.595
SVM	0.864	0.810	0.726	0.396	0.846	0.750	0.821	0.544
LightGBM	**0.967**	0.814	**0.974**	**0.887**	**0.971**	**0.965**	**0.962**	**0.845**

4.2.4 Construction and optimization of the LightCUD method

After determining LightGBM as the core machine learning algorithm, we then present the construction details of LightCUD. For the two 16S rRNA modules, we trained the model parameters with the taxonomic profiles of 503 genera passing the feature selection. With the genus profiles of 201 healthy controls and 148 IBD patients, we trained 16S-based health vs IBD module. With the genus profiles of 127 UC and 21 CD, we trained 16S-based UC vs CD module. The model parameters were tuned to optimize the performance of the two cases through five-fold cross validation. The relatively short reference sequences of the 16S rRNA reference database allowed for the rapid alignment of query sequences, so the 16S modules could work very fast and make judgment for one sample in a few minutes.

For WGS data, we constructed and optimized the WGS-based modules with strains as features. Compared with genera, strains are remarkably more critical for medical interest in judgement of pathogenicity and characterization of a disease. The three-step feature selection strategy evidently reduced the number of features/strains. However, the database including reference genomes of hundreds of strains was still too large for application, so we further shrunk the feature set using a pre-training procedure. At the same time, this procedure could improve the performance and stability of the WGS-based modules.

Through pre-training, we assigned significance scores to the features/strains that passed the three-step feature selection procedure. The significance score of each strain was calculated using LightGBM through evaluating the increment of the error rate caused by removing that strain from the set of predictors. In order to optimize the generalization ability, the size of the feature set was determined and the most important features were selected as follows. All the features/strains were sorted in a queue according to their significance scores. Herein we added the top one feature into a waiting list, and evaluated AUC of the models using the feature in the waiting list with a five-fold cross validation. Then, we added the next one feature into the waiting list and calculated the AUC again. This process was continued until all the features were added into the waiting list, overall 320 strains for health vs IBD case and 159 strains for UC vs CD case. We found that with increasing number of features added into the waiting list, the AUCs increased at the beginning and decreased after reaching the largest values. The WGS-based health vs IBD module with 49 features achieved the best performance, and the UC vs CD module with 12 features achieved the best performance (Figure 4.2). The module AUCs labeled as stars in Figure 4.2 were higher than those of modules shown in Table 4.1, even though the number of feature/strains are reduced, which illustrated that the model performances was improved by optimizing feature set, owing to the reduction of potential noise features. Therefore, for WGS-based modules, we selected the 49 most important features/strains for health vs IBD discrimination and the 12 strains for UC vs CD discrimination. With these features, we separately trained the

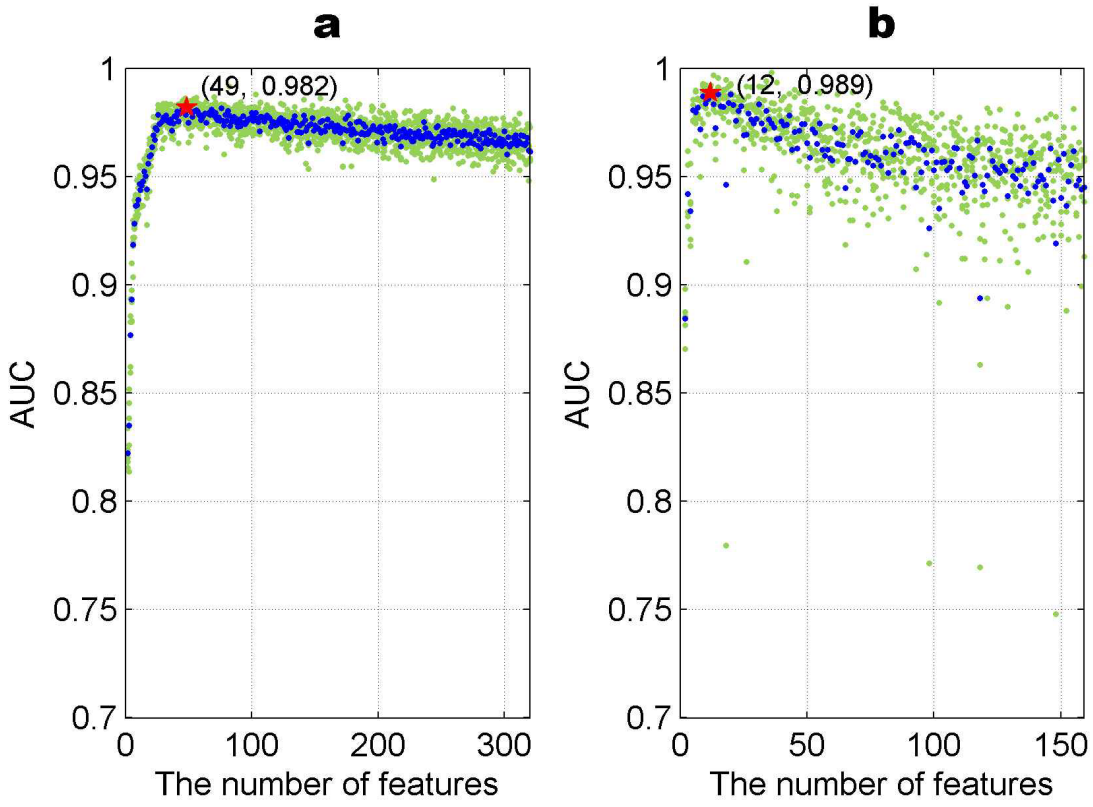

Figure 4.2: Optimizing the feature sets for WGS-based modules. The cyan dots denoted the AUC values for a different round of five-fold cross validation with the different number of features, and the blue dots represented mean values of the cyan dots in the same column. (**a**) Illustrated the WGS-based health vs IBD case and (**b**) illustrated the WGS-based UC vs CD case. For both the cases, AUCs increased with more features at the beginning and decreased after reaching the top values. 49 features for WGS-based health vs IBD case and 12 features for UC vs CD case were best.

health vs IBD module and UC vs CD module.

4.3 Results

4.3.1 Implementation and performance of LightCUD

As the feature sets (genera for 16S-based modules and strains for WGS-based modules) and model parameters were determined, we trained the models of LightCUD with all the training samples to build a universal decision tree. Finally, we released the LightCUD

program for first identifying IBD colitis samples and further discriminating UC and CD, for both WGS data and 16S sequencing data from the human gut microbiota samples. As shown in Figure 4.3, LightCUD goes through different processing routes for WGS data and 16S sequencing data. For WGS data, LightCUD first blasts the raw data in FASTA format against the customized health vs IBD reference database embodying the reference genomes of the 49 strains we determined above. With the alignment results, LightCUD calculates the taxonomic profiles and then determines whether the query sample tends to be healthy or IBD. If IBD is indicated, LightCUD further decides whether the query sample belongs to UC or CD type in this sample using the customized WGS reference database of the 12 selected strains. For 16S data, the genera serving as model features were consistent for both the health vs IBD and UC vs CD modules, so only one 16S rRNA reference database of 503 genera was embodied for these two modules. With 16S rRNA data, LightCUD goes through one time of alignment against the reference database to reveal the taxonomic profile of the query sample on genus level. Further, LightCUD makes decision based on this taxonomic profile in two steps. Firstly, LightCUD decides whether this sample indicates health or IBD. If IBD, LightCUD further makes decision about the specific type of IBD as UC or CD. These four modules have been integrated into an accessible pipeline and released as an open source tool on the webpage (see http://cqb.pku.edu.cn/ZhuLab/LightCUD/), along with the customized databases. The databases were built using the reference sequences of strains from NCBI[95] for WGS-based modules and genera from RDP[134] for 16S-based modules.

The performance of LightCUD was validated using the average AUC and AP with five-fold cross validation. The average AUC and AP of both the WGS-based and the 16S-based modules, for health vs IBD and UC vs CD cases, indicated that all four cases were highly discriminative in distinguishing IBD from healthy controls, and further identifying the specific type of IBD (4.4). It is also noted that the WGS-based modules (AUC = 0.984 and AP = 0.947 for health vs IBD module, AUC = 0.989 and AP = 0.953 for UC vs CD module)

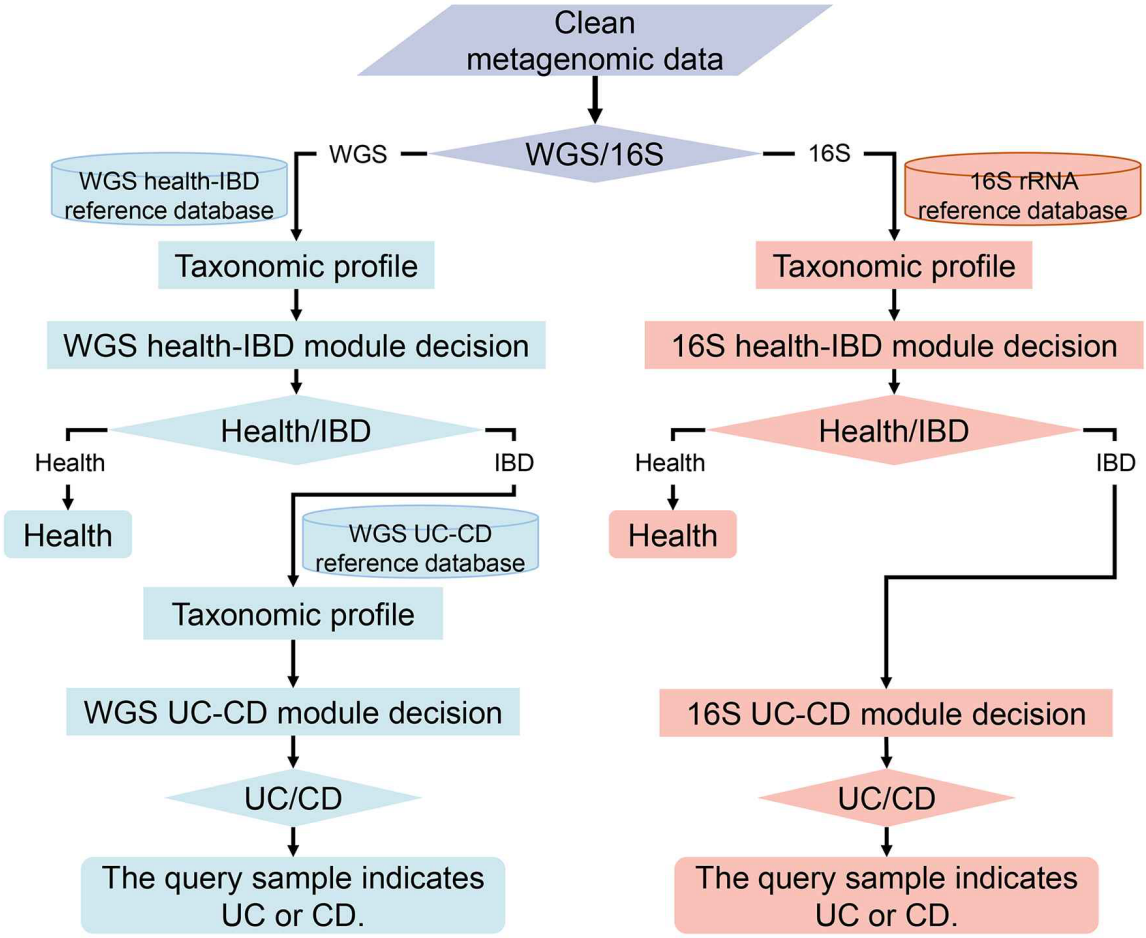

Figure 4.3: Schematic of the LightCUD framework. The input data to LightCUD is the raw reads of the sample in FASTA format. First, with the '-t' parameter, LightCUD decides the data type. For different data types, different customized reference databases are used. For both WGS and 16S data, LightCUD goes through a two-stage judgment. At the first stage, LightCUD decides whether the query sample is healthy or IBD. If IBD, LightCUD further judges the specific type, namely, UC or CD.

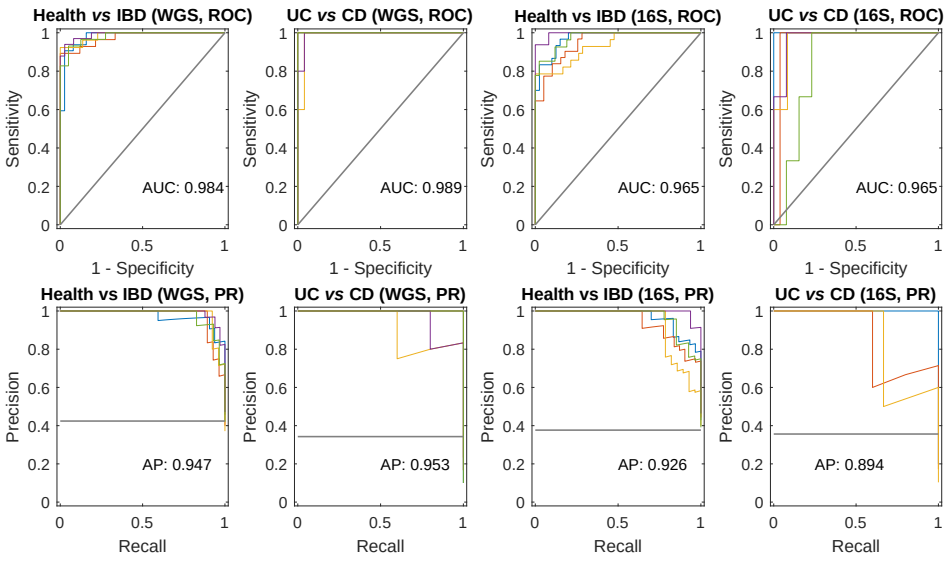

Figure 4.4: Evaluation of the performance of LightCUD. We evaluated the accuracy of disease classification using LightCUD with receiver operating characteristic curve and precision-recall curves representing the results. Lines in each subplot with different colors represent the model performance in one of the five-fold cross validations. As the training sets were unbalanced, we reported both the AUC values and the AP values. The average AUC and AP were labeled under corresponding curves.

performed better than the 16S-based modules (AUC = 0.968 and AP = 0.926 for health vs IBD module, AUC = 0.965 and AP = 0.894 for UC vs CD module). For the current release, we set default discrimination thresholds with regard to the sample proportion of training data. For health vs IBD cases, default thresholds were set as $N_{IBD}/(N_{IBD}+N_{Health}) = 0.42$, wherein N represents the number of samples in corresponding class labeled with the subscripts. Similarly, default thresholds were set as $N_{CD}/(N_{CD}+N_{UC}) = 0.14$ for UC vs CD cases. With the default thresholds, LightCUD reached high prediction accuracies during five-fold cross validation, on the average, 92.3% for WGS-based health vs IBD module, 93.3% for WGS-based UC vs CD module, 88.5% for 16S-based health vs IBD module, and 93.1% for 16S-based UC vs CD module. Herein, the accuracies were the proportion of predicted labels that were exactly the same as the actual labels of samples.

In order to verify the generalization ability, we further conducted blind validation with

an independent dataset with 185 samples including 16 healthy controls and 169 with CD [16]. After removing low quality reads and human genome reads, we run LightCUD on the sequences of each sample. The program returned a score indicating IBD probability. The results showed that LightCUD maintained good performance (AUC = 0.809, AP = 0.971) in discriminating healthy controls from IBD patients with CD. Further, LightCUD showed 76.9% accuracy when discriminating CD from UC.

4.3.2 The strains for WGS modules serving as biomarkers

The high performance of the strain-level WGS-based module convinced us that the 49 strains discriminating healthy controls from IBD patients and the 12 strains distinguishing the specific type of IBD were clinically valuable. More details about these strains as biomarkers are presented as follow.

The 49 strains selected for health vs IBD discrimination, and the 12 strains for UC vs CD discrimination were completely different, as shown in Figure4.5c and Figure4.5d. It should be noted that most of the selected disease sensitive strains were not dominant members in the microbial community. In Figure4.5a and Figure4.5b, the features passing the three-step selection process were sorted in a descending order according to feature significance scores assigned during pre-training. The relative abundance of each feature was randomly distributed. This observation excluded the possibility that the features we finally selected were in extremely low abundances, also indicated that the features with valuable discrimination ability are not necessarily dominant.

These strains serving as discriminative biomarkers would be valuable for the analysis of IBD related intestinal microbial dysbiosis. We discussed some reported findings associated with these biomarkers. *Enterobacter cloacae*, a clinically significant species, has been reported to be enriched in the intestines of IBD patients[186]. As a member of this species, in our study strain *E. cloacae* subsp. dissolvens SDM was significantly enriched in IBD samples and very important for the health vs IBD discrimination (Figure4.5c). In addition,

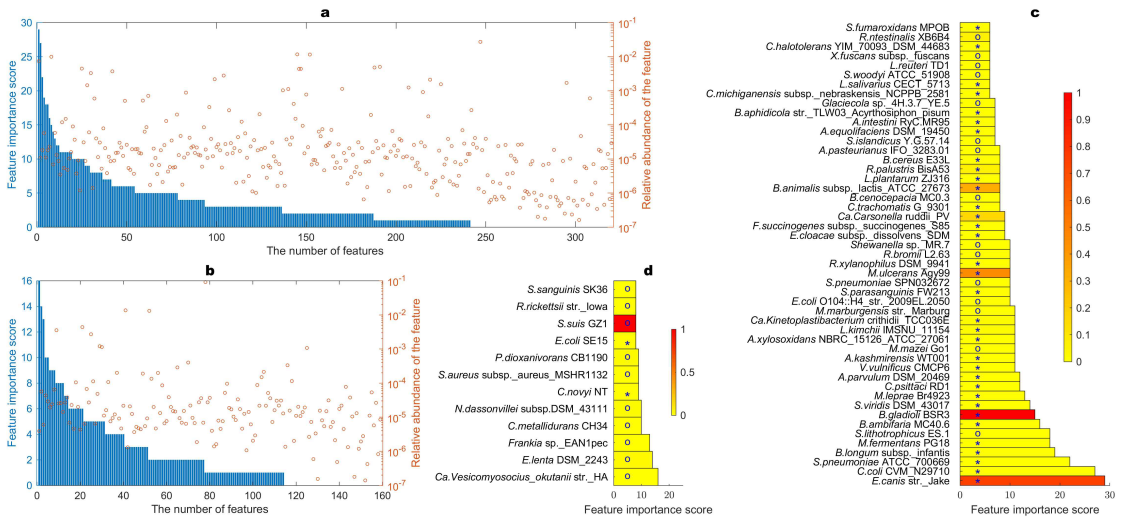

Figure 4.5: Features abundances in light of feature significance scores. Relative abundances of features for the WGS-based health vs IBD module (**a**) and the UC vs CD module (**b**). All the features that passed the three-step feature selection were shown in descending order according to feature significance score. (**c**, **d**) Color bars show relative abundances of features, scaled to 0-1 with the maximum value of all 30 (or 15) abundances values. In (**c**), '*' indicates significantly higher abundances in IBD and 'o' indicates the abundances of strains significantly decreased in IBD (P¡0.01). In (**d**), '*' indicates significantly higher abundances in CD and 'o' indicates significantly higher abundances in UC (P¡0.01).

species *Mycobacterium ulcerans* has been reported to be the major cause of the skin disease Buruli ulcer[187], herein the *M. ulcerans* str. Agy99 was enriched in IBD fecal samples and exhibited quite high discrimination ability. Furthermore, species *Burkholderia gladioli* and *Ehrlichia canis* have both been frequently reported as pathogens[188, 189], in this study the *B. gladioli* str. BSR3 and *E. canis* str. Jake with high discrimination ability were enriched in IBD samples. *Nocardiopsis dassonvillei* subsp. dassonvillei DSM 43111 has been reported to produce cellulases when provided with appropriate substrates, and the genome also has sequences for six predicted cellulose degrading enzymes, which are necessary for digesting fiber and cannot be produced within the body[190]. Cellulose has been proved effective for colitis amelioration[191]. Our results revealed that this strain was quite important for UC and CD discrimination, and was significantly depleted in CD samples compared with UC samples.

4.4 Discussion

In this study, we constructed a diagnostic tool, LightCUD, which can discriminate IBD from healthy controls and further distinguish the specific type of IBD. The LightCUD program performed well for both WGS and 16S sequencing data, AUC ¿ 0.95 and AP ¿ 0.89 for all four cases (WGS-based health vs IBD and UC vs CD, 16S-based health vs IBD and UC vs CD) during five-fold cross validation. As the first released human gut microbiome-based IBD diagnostic tool, LightCUD embodies discrimination modules constructed with stool samples better than any other reported classifiers. Gevers et al. constructed a classifier to distinguish CD from healthy controls based on 199 stool samples with an AUC of 0.66 [8]. Papa et al. performed an analysis of the 16S sequencing data from 91 stool samples, and reached a discrimination accuracy with an AUC of 0.83[192]. For the current study, LightCUD has only a single command line but provides a non-invasive mechanism of distinguishing IBD from healthy controls based on stool samples. For either WGS or 16S data, LightCUD processes a sample in several hours, depending on the sequencing depth. Parallel computation may further reduce the running time of prediction. With the development and popularity of NGS, LightCUD highlights the potential of diagnostic tools developed with machine learning algorithms based on the data of human gut microbiome.

It should be pointed out that the WGS-based module construction involved a well-designed selection of features, which contributed to a set of highly representative features (strains) of the microbial community and accelerated the computation. The analysis of these selected features revealed the fact that highly discriminative features were not necessarily to be dominant ones. In addition, these features/strains have been frequently reported as IBD associated strains. These specific strains are valuable biomarkers in designing animal models of IBD for human clinical trials to study the mechanisms of probiotics and pathogens in ameliorating inflammation.

LightCUD may be subject to the available sample sets. For example, the model dis-

tinguishing UC from CD performed inferiorly than the model distinguishing IBD from healthy controls, because of the limited number of CD samples in our training set. Also, LightCUD was subject to limited clinical trials. Therefore, even though LightCUD outperformed the other reported programs, further validation is essential before it can be used in clinical diagnosis.

CHAPTER 5
A DATABASE INTEGRATING DISEASE-RELATED MARKER GENES IN HUMAN GUT MICROBIOME

5.1 Introduction

In the past decade, the correlation between gut flora and the pathological disorders of host has become the hotspot of research, remarkably further driven by the launch of the National Microbiome Initiative (NMI) in 2016. The clinical trials and animal experiments have revealed that the alteration of the gut microbiota are closely associated with the happen of diseases like type II diabetes (T2D)[32, 33], Crohn's diseases (CD)[34], obesity[35], depression[36] or colorectal cancer[37], illustrated by obvious changes in community structure and metabolic potential. For example, it has been proved that the relative portion of *Bacteroidetes* decreased in obese people, while increased with weight loss on two types of low-calorie diet[40]. Gut microbiota from inflammatory bowel disease (IBD) patients were detected to produce significantly more short-chain fatty acids and ammonia than that from healthy individuals[41]. Also, studies of depression have demonstrated the overrepresentation of *Bacteroidales* and underrepresentation of *Lachnospiraceae*[42]. These studies indicated that the gut flora are strongly associated with human health generically and biochemically, therefore led to a stirring of interests of variation in the gut microbiota at both species and gene level.

In regard to all this progress, the rapid development of metagenomics using next-generation sequencing (NGS) has been playing an important role in understanding the impact of gut microbial ecosystem on human disease. Through case-control metagenome-wide association studies, the population structures of gut flora have been well studied. IBD-affected individuals were reported to have 30-50% reduced biodiversity of gut mi-

crobes[38]. Studies focusing on bacterial 16S rRNA gene phylotypes suggested significant phylotype level alterations in the intestinal microbiota of irritable bowel syndrome (IBS) patients[36, 39]. However, 16S rRNA analyses have limitations on understanding the microbial molecular-mechanisms. Thus the whole-genome sequencing (WGS) has been widely used to investigate host microbial metabolic potentials. An obesity-related study demonstrated that the obesity-associated signals originating from the host gut microbiome may be much stronger than that of presently known host genome[35]. A T2D-related study showed that the T2D-enriched microbial markers involve membrane transport of sugars, branched-chain amino acid transport, methane metabolism, xenobiotics degradation and metabolism, and sulphate reduction[33]. It is also documented that from the point of gut flora, liver-cirrhosis was associated with assimilation or dissimilation of nitrate to or from ammonia, GABA (c amino butyric acid) biosynthesis, denitrification, GABA shunt, phosphor transferase systems, haem biosynthesis and some types of membrane transport, such as amino-acid transport[43]. These pathogenic studies of gut microbiome have exhibited a lot of disease-related genes and metabolic pathways of gut flora involved in those kinds of known human diseases. For instance, 15,894 genes were indicated as the significantly different genes, they may not only be applied to an efficient discrimination of lean and obese individuals, but also characterize human gut flora by gene functionary level for obesity-related metabolic disorders[35]. For gut microbial species associated with liver cirrhosis, there were 75,245 genes identified to reveal the difference between patients with liver cirrhosis and health controls[43]. Since these disease-related genes are disease specific and population specific, it should be expected that more batches of such genes related with human diseases will be found in gut microbiome with the expanding of human microbiome projects.

Although some a set of microbial genes may be recognized as those associated with a specific disease, we probably just see the forest for the trees without a common set of microbial genes related to human health and diseases in general terms. People constructed

such as the IGC database, an integrated gene catalogue of intestinal microbiome, however it is not associated with specific pathogenicity and host health[66]. Therefore, an integrated general knowledge of these microbial genes certainly facilitate understanding the correlation between gut flora and human health and disease, as well as the mechanisms of how gut flora contributes to disease process. However, at present large amounts of metagenomic data by current studies widely scatter at the public databases GenBank[95] and EMBL[96], leading to the difficulty in information integration and utilization, as well as to a provisional knowledge with fragmentary evidence.

Herein we constructed a comprehensive database, named DREEM, of Disease-RElatEd Marker genes in human gut microbiome, which have retrieved a large scale of WGS data of human gut metagenomes in DREEM, covering six types of pathological conditions, i.e., T2D, CD, ulcerative colitis (UC), liver cirrhosis, symptomatic atherosclerosis and obesity. The short reads with the size of 18.63T consisting of 1,729 samples were processed with a standard procedure, involving the state-of-the-art bioinformatics tools and statistical analysis. Then we picked out 1,953,046 non-redundant DREEM genes. The DREEM genes specific to a certain disease were also stored as an individual gene set with respect to six diseases considered in the current program. Furthermore, we provided a set of Core-DREEM genes, which are shared among the samples of five metabolic syndrome related co-morbidities: T2D, CD, UC, liver cirrhosis and obesity. All DREEM genes were analyzed for taxonomic classification and functional annotation. Serving as the integration of gut microbial pathogenic gene catalogues, as a result, DREEM could be employed to detect functional and metabolic disturbance of host gut microbiomes, thus may provide brand-new strategies for host disease diagnosis and facilitate studies on human gut flora.

5.2 Data and methods

5.2.1 Data source

The data in DREEM were collected from seven studies[33, 32, 43, 35, 78, 79, 80] currently with all WGS metagenomics data publicly available, which include samples with six diseases (T2D, obesity, CD, UC, liver cirrhosis, symptomatic atherosclerosis, Table 5.1) and their corresponding control groups. We collected 18.63T paired-end short read sequences of 1,729 samples from GenBank and EMBL (Table 5.2).

Table 5.1: Original publications of samples

Diseases	Major References
T2D	[1]Qin, J., et al. (2012). Nature, 490(7418): 55-60.
T2D	[2]Karlsson F.H., et al. (2013). Nature, 498(7452): 99-103.
CD, UC and Obesity	[3]Nielsen H.B., et al. (2014). Nature biotechnology, 32(8): 822-828.
CD	[4]Lewis, J.D., et al. (2015). Cell Host & Microbe, 18(4): p. 489-500.
Obesity	[5]Le Chatelier, E., et al.Nature, 2013. 500(7464): p. 541-546.
Liver Cirrhosis	[6]Qin N.,et al. (2014). Nature, 50(2):311-317.
Symptomatic Atherosclerosis	[7]Karlsson F.H., et al. (2012). Nature communications, 3: 1245.

Table 5.2: Detailed information of samples and DREEM genes

Disease Type	Soure Paper	Sample NUM	Control	Patients	Num of DREEM Genes
T2D	[1]	352	173	179	331,420
	[2]	96	43	53	
CD	[3]	223	201	21	660,972
	[4]	369	26	343	
UC	[3]	328	201	127	748,605
Obesity	[3]	274	163	111	469,580
	[5]	265	96	169	
LC	[6]	237	114	123	355,170
Atheor	[7]	27	15	12	21,730

5.2.2 Data processing

All the WGS sequencing data were retrieved and processed in a standard workflow shown in Figure 5.1. We now describe in detail the steps of data processing as follows.

Short reads assembly and gene annotation

With the original short reads assembled into contigs by InteMAP, an integrated metagenomic assembly pipeline designed for NGS short reads[183], genes were then annotated by two metagenomic gene predictors, MetaGeneMark[193], MetaGUN/MetaTISA[194, 103], with the strategy that genes detected by the two tools were combined to include more protein coding genes. Short reads with low quality and contigs with low sequencing depth were strained off for the validity of further analysis.

Significant genes identification and their validation

The annotated genes from each of the seven studies[33, 32, 43, 35, 78, 79, 80] were first respectively clustered by CD-HIT[104], a standardized and ultra-fast clustering algorithm used to merge similar gene sequences, and seven sets of non-redundant genes selected by CD-HIT were then obtained. Furthermore, implemented via the alignment tool bowtie2[195], the occurrence of each non-redundant gene was calculated by the number of mapped reads. The occurrences of genes then resulted in an abundance matrix, serving for further calculation of significant difference. Here Wilcoxon rank sum test is applied. As a nonparametric test approach, it found out genes with significantly different distribution between two sample groups using magnitude-based ranks. If the ranks of the two sample groups are significantly separated, then the statistic test identifies significant difference[196]. In this paper, the genes with significantly higher frequencies ($P < 0.05$ in Wilcoxon rank-sum test, which is a widely accepted significant level) in disease group than in healthy control group were picked out as the significant genes.

To show the stability of significant genes identified with respect to the sample size of

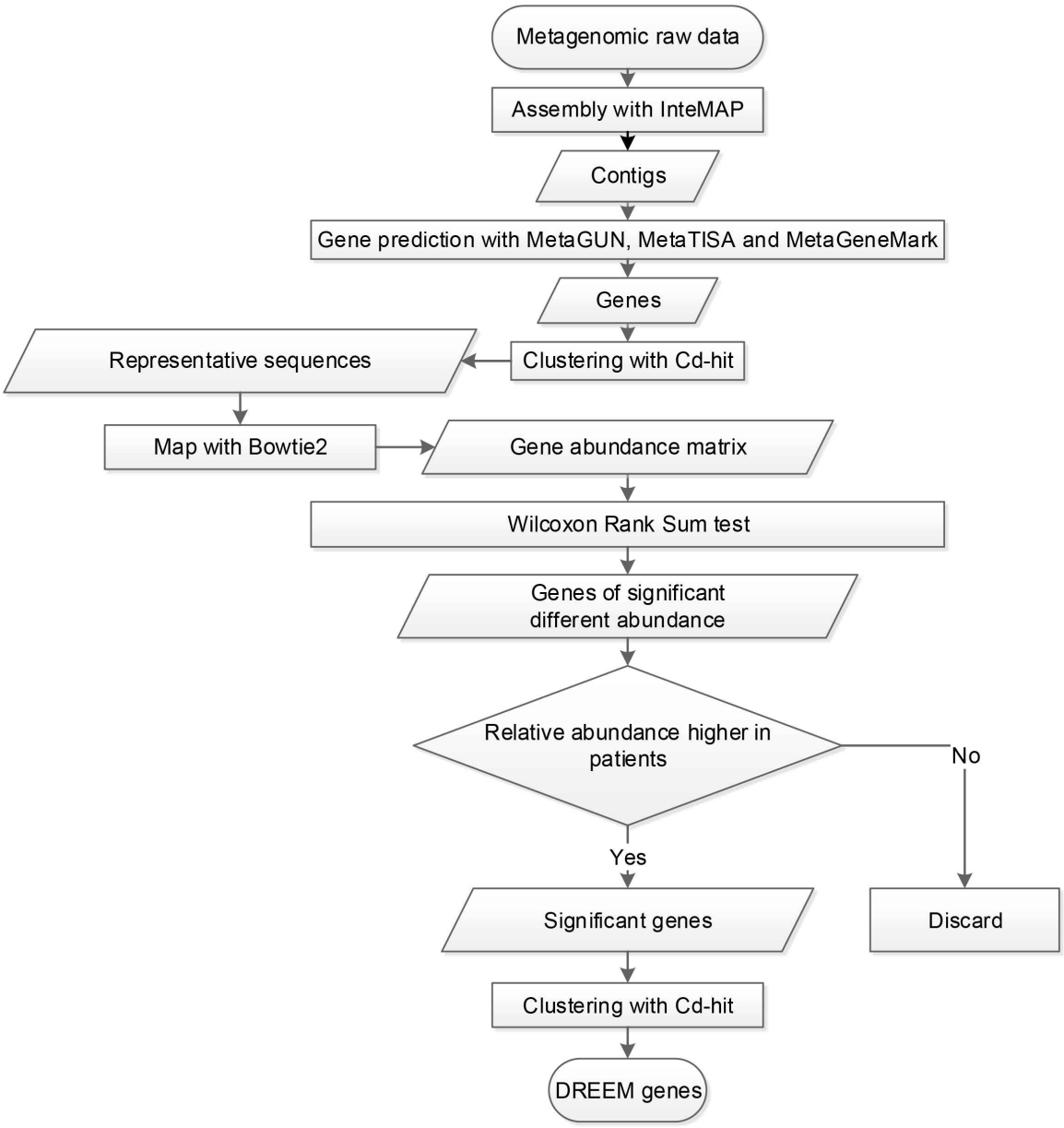

Figure 5.1: Data processing workflow. All the WGS data downloaded from GenBank or EMBL are processed in this standard procedure. Original WGS data are assembled into contigs by InteMAP, from which genes were further predicted by MetaGUN and MetaTIS-D. Gene sequences were clustered with CD-HIT (sequence identity threshold set at 90%) to generate presentative sequences. In the unit of publication, Significant genes were selected after W-rank sum test ($P < 0.05$). Finally, all significant genes were clustered again to reduce redundancy and generate the final DREEM genes.

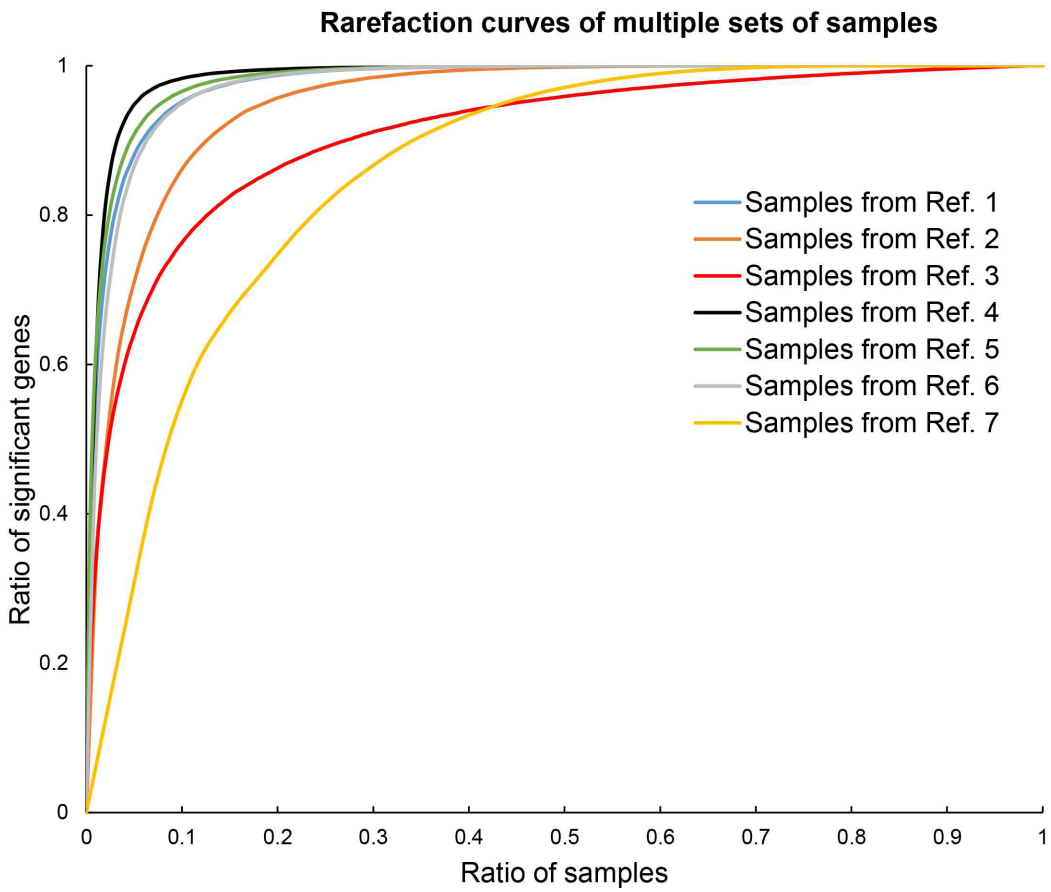

Figure 5.2: Rarefaction curves of the number of significant genes as a function of the number of samples from each publication. The number of samples and significant genes were both normalized by their maximum value to fit to one graph. All the AUC values exceeded 0.85, which indicated that the sample number is sufficient for almost all the potential significant genes related to the six types of diseases.

seven studies used by us, we also conducted a saturation evaluation for each of seven sets of significant genes (Figure 5.2). The number of significant genes is a function of sample numbers. Since all the AUC (area under the receiver operating characteristic curve, see details in Table 5.3) values achieved greater than 0.85, which is large enough to confirm that we effectively identified almost all the genes significantly related to the relevant six diseases.

Table 5.3: AUC of rarefaction curves of number of significant genes as function of sample numbers

	AUC	Lower Boundary	Upper Boundary
Samples from Ref. 1	0.9773	0.9678	0.9838
Samples from Ref. 2	0.9518	0.9203	0.9715
Samples from Ref. 3	0.9096	0.8865	0.9312
Samples from Ref. 4	0.9851	0.9749	0.9909
Samples from Ref. 5	0.9811	0.9747	0.9856
Samples from Ref. 6	0.9742	0.9561	0.9854
Samples from Ref. 7	0.8601	0.7054	0.9401

DREEM genes and their functional annotation

Considering all the seven sets of significant genes mentioned above as a whole, we then integrated the data sets and clustered them with CD-HIT[104] again to further reduce the redundancy. Thus, we retrieved the final representative genes as the DREEM genes (Disease-RElatEd Marker genes) set. A total of 1,953,046 DREEM genes were collected as the final set, which covers the six diseases, and 58.9% of which are complete (with 5' end and 3' end) coding gene sequences.

Gene function prediction was further performed for all the DREEM genes via BLAST[197] against the COG database[198] (e-value$\leq 10^{-5}$).

5.3 Results

5.3.1 Data organization and statistics

As mentioned above, a total of 1,953,046 DREEM genes was collected for the DREEM dataset. Furthermore, they were classified into six groups as DREEM_T2D, DREEM_Obesity, DREEM_Crohn, DREEM_UC, DREEM_LC, and DREEM_Athero, corresponding to each disease. Different diseases (e.g., T2D and liver cirrhosis, liver cirrhosis and IBD, IBD and T2D) show a relatively unique profile, though many DREEM genes were shared among them. Venn diagram was drawn to show overlaps among the five sets of DREEM genes, i.e.,

DREEM_T2D, DREEM_Obesity, DREEM_Crohn, DREEM_UC and DREEM_LC (Figure 5.3, Table 5.4). The gene set DREEM_Athero was not included owing to there was less correlation of Atherosclerosis with gut microbial community, compared to the other five types of diseases, indicated by the extremly small number of DREEM genes in the data set of DREEM_Athero. Herein, 5,100 DREEM genes were found shared by the five types of diseases, and were designated as the Core-DREEM genes. For public releasement, the sequences in the set of Core-DREEM genes were uniformly renumbered (see detailed information on the database website).

We applied PhymmBL[185] to taxonomically classify the DREEM genes, 75% of which were assigned into 14 phylums and 581 species. Overall, for both the DREEM genes and the Core-DREEM genes, the dorminant phylums are Proteobacteria, Firmicutes and Bacteroidetes, and the most abundant species are *Coprococcus catus* GD/H7, *Bacteroides fragilis* NCTC9343, *Bacillus thuringiensis* YBT1518, *Clostridium perfringens* str.13 (details are illustrated on the DREEM database website, Table 5.5, Figure 5.4 and Figure 5.5).

Table 5.5: Number of core DREEM genes assigned to each species (only numbers > 20 were shown)

Species	NUM
Coprococcus_catus_GDSLASH7	835
Bacteroides_fragilis_NCTC_9343	435
Bacillus_thuringiensis_YBT-1518	198
Clostridium_perfringens_str._13	118
Desulfomicrobium_baculatum_DSM_4028	112
Marinobacter_hydrocarbonoclasticus_ATCC_49840	97
Eggerthella_sp._YY7918	79
Cronobacter_sakazakii_SP291	75
Bifidobacterium_adolescentis_ATCC_15703	70

Table 5.5 continued

Desulfatibacillum_alkenivorans_AK-01	55
Escherichia_coli_O7COLONCOLONK1_str._CE10	55
Enterobacter_cloacae_subsp._cloacae_ENHKU01	51
Treponema_azotonutricium_ZAS-9	51
Staphylococcus_saprophyticus_subsp._saprophyticus_ATCC_15305	50
Carnobacterium_sp._17-4	49
Desulfobulbus_propionicus_DSM_2032	49
Ilyobacter_polytropus_DSM_2926	45
Lactococcus_lactis_subsp._cremoris_A76	44
Yersinia_pestis_D182038	40
Escherichia_coli_O83COLONCOLONH1_str._NRG_857C	37
Rhizobium_leguminosarum_bv._viciae_3841	37
Nitrosococcus_watsonii_C-113	36
Desulfosporosinus_orientis_DSM_765	34
Pandoraea_sp._RB-44	34
Salmonella_enterica_subsp._enterica_serovar_Choleraesuis_str._SC-B67	34
Escherichia_coli_042	33
Herbaspirillum_seropedicae_SmR1	32
Buchnera_aphidicola_str._Ak_LPARENAcyrthosiphon_kondoiRPAREN	31
Thermaerobacter_marianensis_DSM_12885	31
Acidovorax_avenae_subsp._avenae_ATCC_19860	29
Flavobacterium_indicum_GPTSA100-9_=_DSM_17447	29
Streptococcus_agalactiae_A909	29
Streptococcus_agalactiae_NEM316	29
Candidatus_Nitrospira_defluvii	28
Streptococcus_suis_GZ1	28

Table 5.5 continued

Azospirillum_brasilense_Sp245	27
Streptococcus_pneumoniae_gamPNI0373	25
Bartonella_clarridgeiae_73	24
Methanocella_arvoryzae_MRE50	24
Nitrosomonas_sp._AL212	24
Thermoplasma_acidophilum_DSM_1728	24
Dickeya_dadantii_3937	23
Pseudomonas_putida_NBRC_14164	23
Bacillus_cereus_B4264	22
Burkholderia_sp._YI23	22
Hyphomonas_neptunium_ATCC_15444	22
Solibacillus_silvestris_StLB046	21

Furthermore, most of the DREEM genes were functionally classified into three main categories, i.e., metabolism, cellular processes and cell signaling. For Core-DREEM genes, 4,805 genes out of 5,100 were conferred with certain functions, among which about 41.1% were responsible for metabolism, while 27.8% for information storage and procession, and 23.1% for cellular processes and cell signaling (Table 5.6, Figure 5.6 and Figure 5.7). Intensively involved pathways include integrase, type IV secretory pathway, site-specific DNA recombinase related to the DNA invertase Pin, DNA-binding response regulator and ABC-type multidrug transport system, ATPase components of ABC transporters with duplicated ATPase domains, Na+-driven multidrug efflux pump, etc.

Table 5.4: Number of DREEM genes shared by different data sets

Data Sets	Number of Genes
T	156556
C	349320
L	212567
O	342846
U	445936
TC	19110
TL	13417
CL	29053
TO	15184
CO	23420
LO	9908
TU	27480
CU	118282
LU	12270
OU	27143
TCL	9757
TCO	5204
TLO	2781
CLO	2746
TCU	28692
TLU	6694
CLU	18704
TOU	4975
COU	13204
LOU	2062
TCLO	1444
TCOU	9397
TLOU	1128
CLOU	3038
TCLU	24501
TCLOU	5100
Total	**1941919**

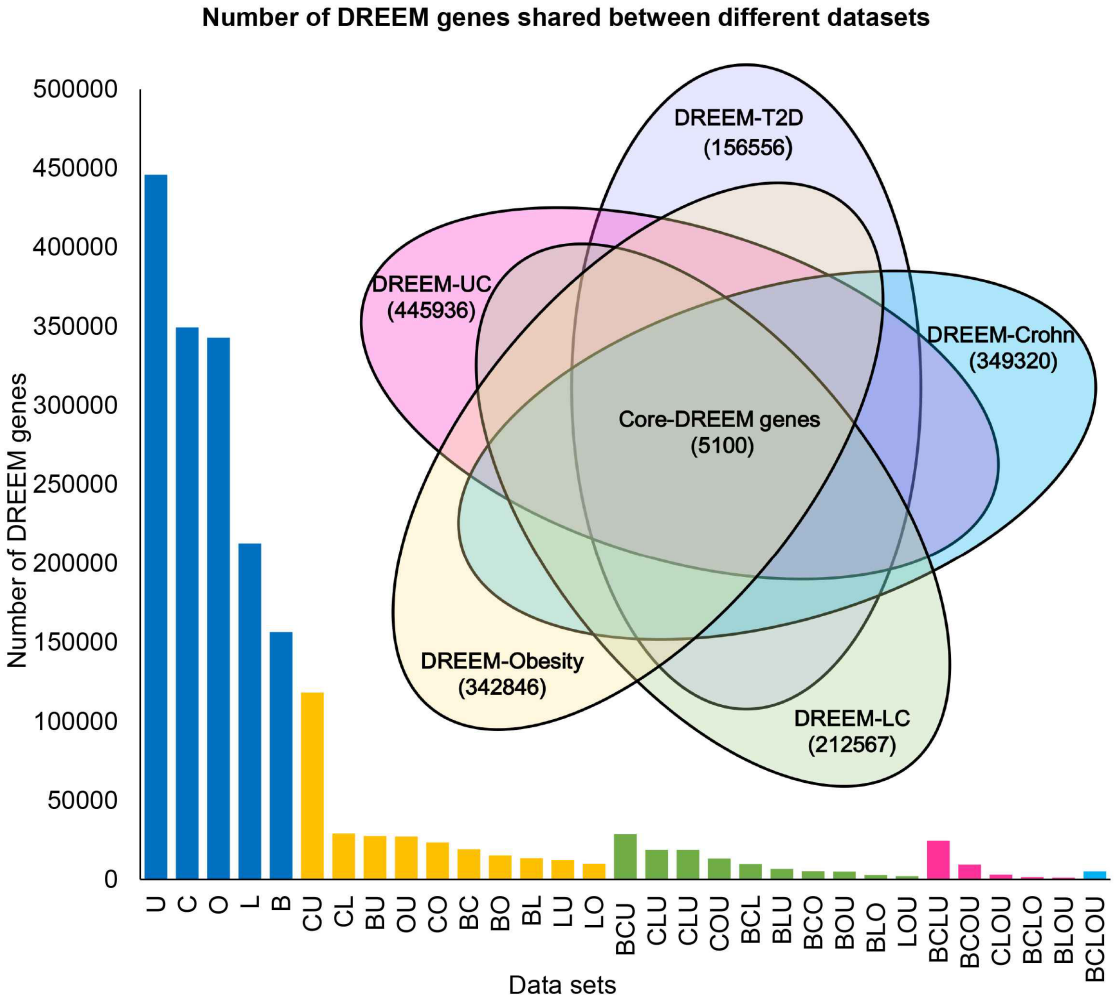

Figure 5.3: Statistics indicating the number of DREEM genes shared by different data sets. U, C, O, L and T stand for data set of DREEM UC genes, DREEM Crohn genes, DREEM Obesity genes, DREEM LC genes and DREEM T2D genes respectively. There are 5,100 Core-DREEM genes, which are shared by other five data sets. Most DREEM genes are unique to one data set. Nevertheless, DREEM UC and DREEM Crohn share the largest number of genes compared with other pairs of data sets, indicating a stronge correlation between the two types of IBD.

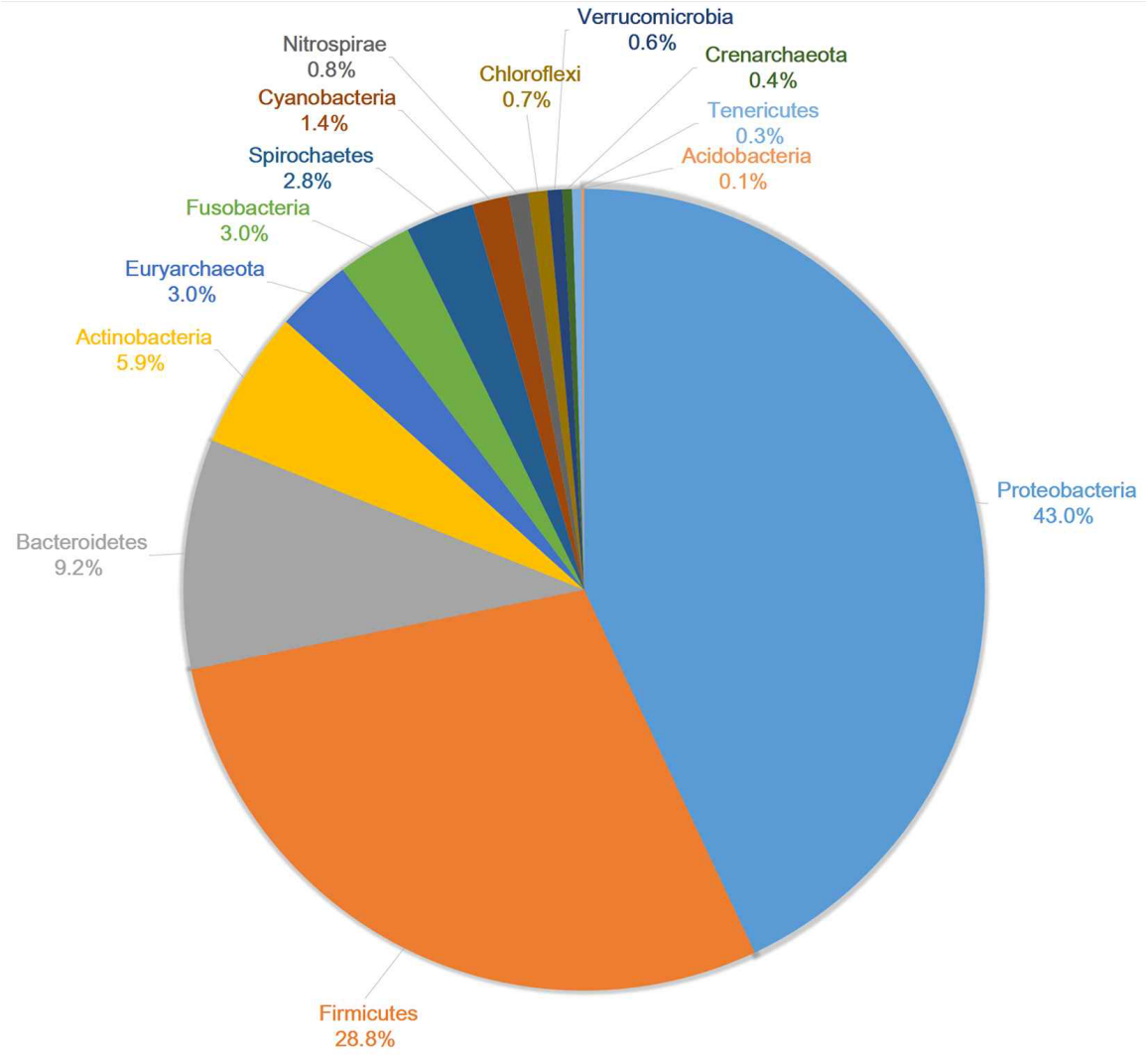

Figure 5.4: Taxonomic annotation of all the DREEM genes at phylum level.

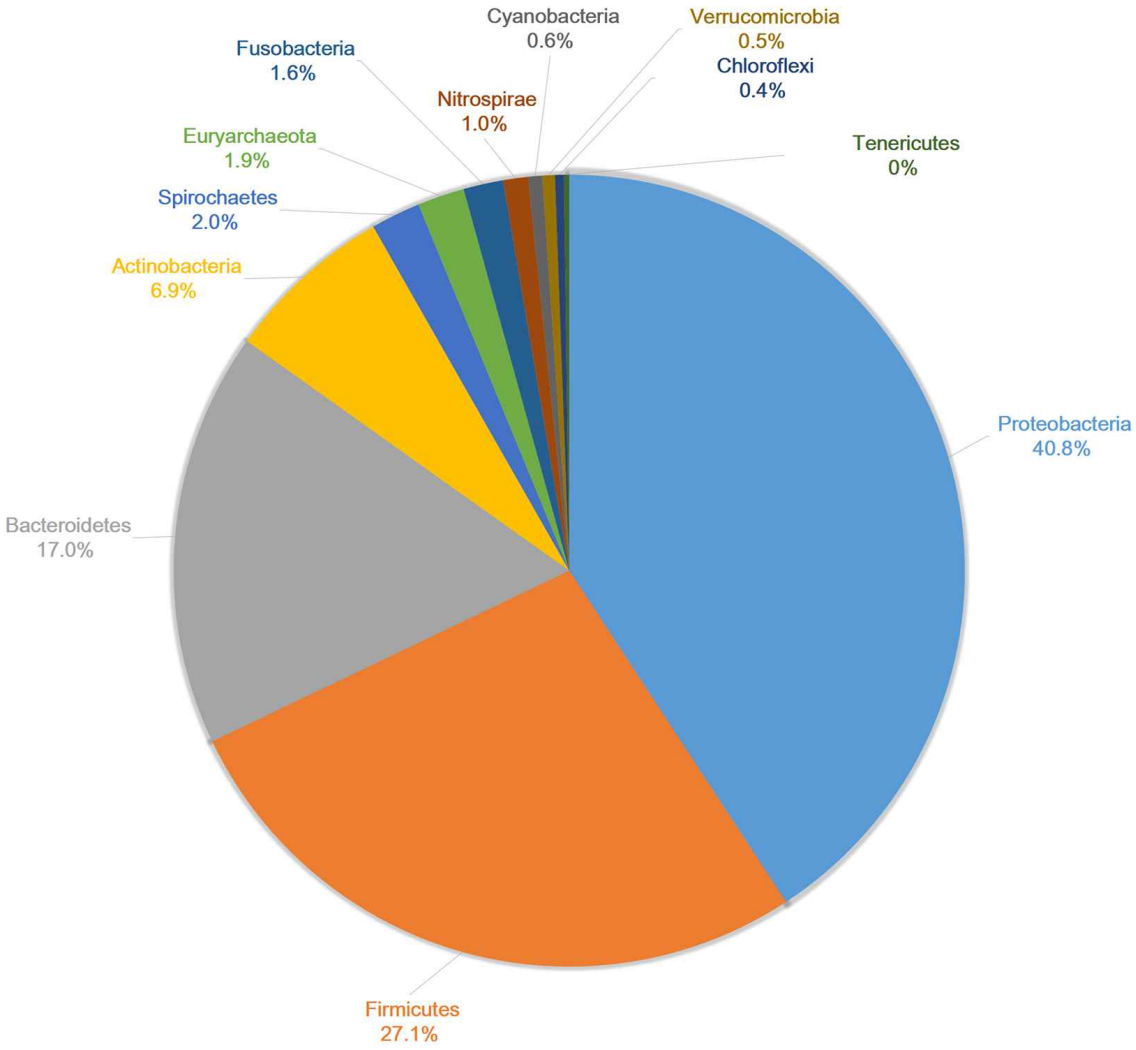

Figure 5.5: Taxonomic annotation of the core DREEM genes at phylum level.

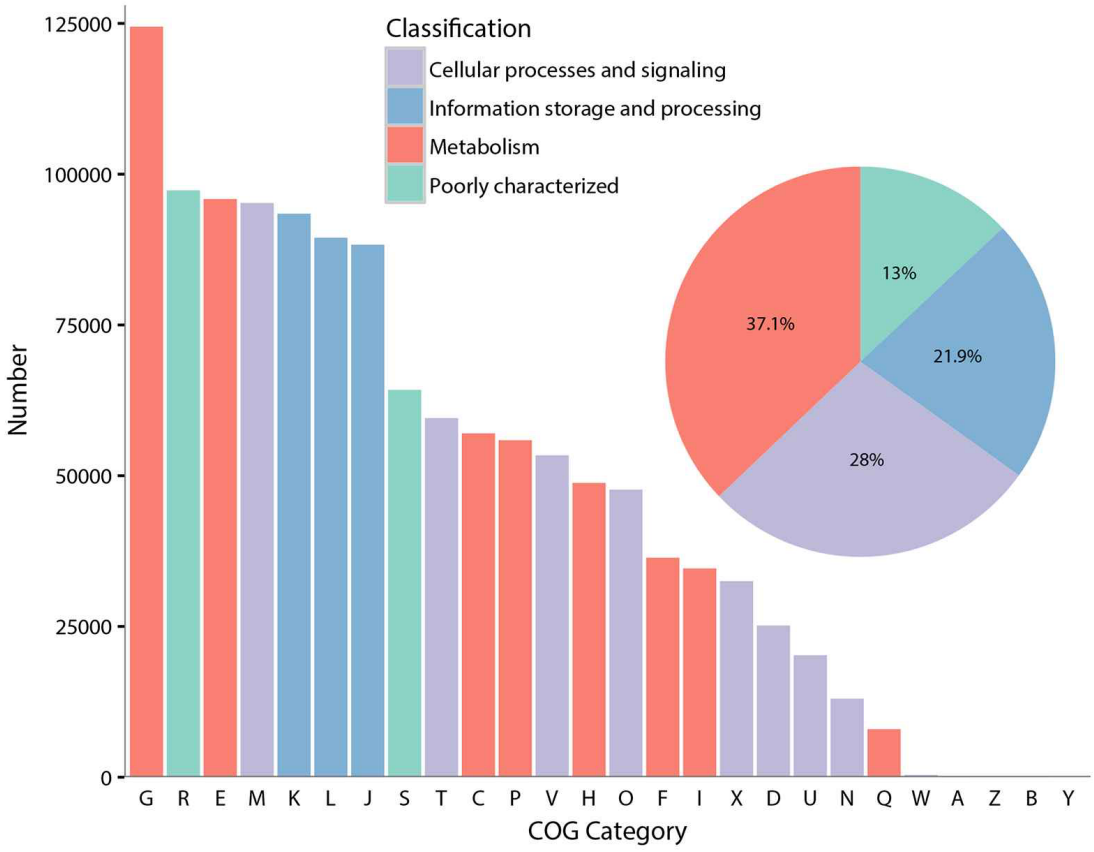

Figure 5.6: Functional annotation of all the DREEM genes via BLAST against COG database ($evalue \leq 10^{-5}$).

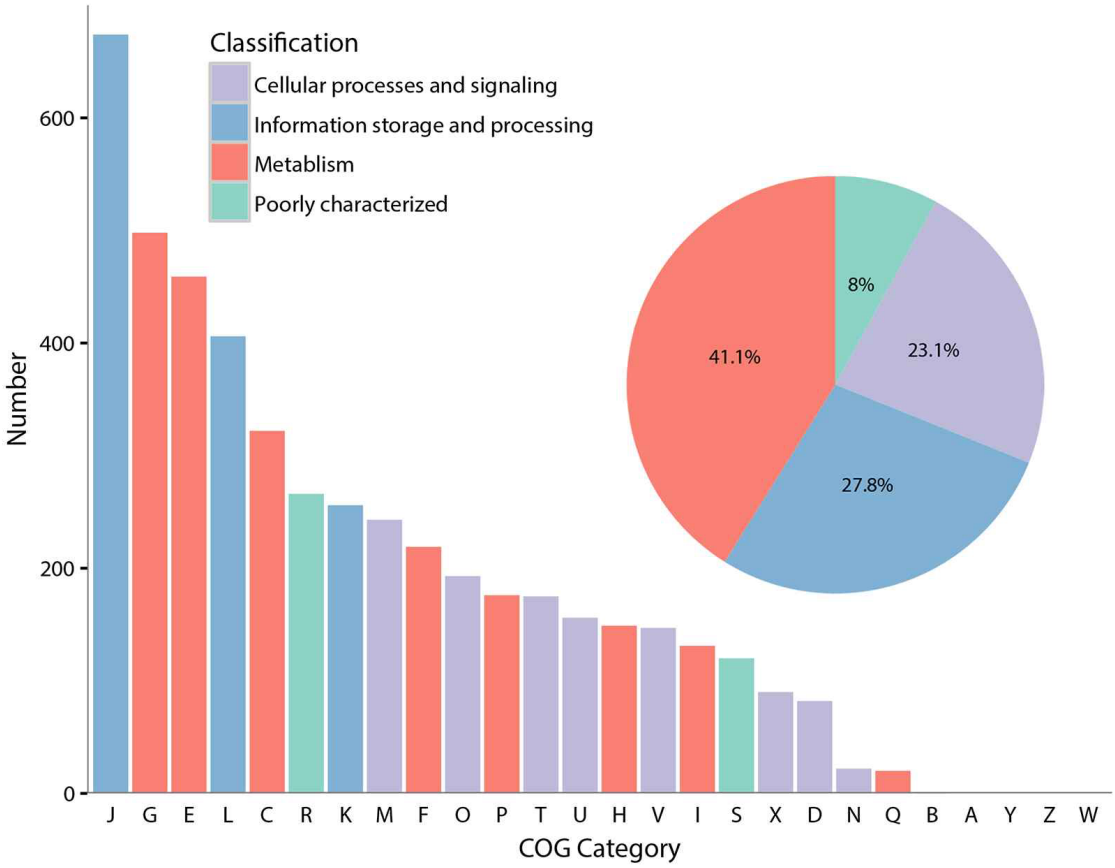

Figure 5.7: Functional annotation of the core DREEM genes

Table 5.6: Number of core DREEM genes assigned to different COGs

	COG category	Number
Information storage and procession	JAKLB	1337
	J	674
	A	0
	K	256
	L	406
	B	1
Cellular processes and signaling cellular	DYVTMNZWUOX	1108
	D	82
	Y	0
	V	147
	T	175
	M	243
	N	22
	Z	0
	W	0
	U	156
	O	193
	X	90
Metabolism	CGEFHIPQ	1974
	C	322
	G	498
	E	459
	F	219
	H	149
	I	131
	P	176
	Q	20
Poorly characterized	RS	386
	R	266
	S	120
Total		4805

5.3.2 Web interface of DREEM

We designed a downloading webpage (http://cqb.pku.edu.cn/ZhuLab/ DREEM/) to allow users to download all the datasets described above. Every gene ID was designed as serial number in the DREEM database, followed by its raw data source information, linked disease categories, integrity description, taxonomic and functional annotation. For those complete genes with complete 3' end and 5' end, length information was also stored in IDs. Every entry in the DREEM dataset is constituted by a gene ID followed by nucleotide sequence. There are eight sets in total on our webpage, one set of all the DREEM genes, six sets of six types of diseases separately, and one set of the Core-DREEM genes. On the search page, one may search for interesting items of the Core-DREEM genes. On each item of search result, apart from taxonomic and functional annotation information, linkage to NCBI is also available for further investigation.

5.4 Discussion

With numerous metagenomic data, there are continuous publications focusing on the impact of gut microbiome on host diseases, from which more and more important disease-related genes and pathways have been discovered. To integrate these genes, DREEM was set up from 18.63T paired-end short reads of 1,729 human gut microbiota samples, and 1,953,046 DREEM genes were constructed through a well-designed procedure. As the first database concentrating the disease-related genes of human gut microbiome, DREEM compiled multiple sets of high-throughput sequencing data of disease-related metagenomic studies including six diseases. Furthermore, the saturation evaluation (as shown in Figure 5.2) for each set of significant genes suggests that the sample size is enough to cover almost all the potential DREEM genes. Another merit of DREEM is the identification of the Core-DREEM genes, which are released as an important set on the webpage. Clearly, based on the DREEM database, the road to an integrated general knowledge may be paved with

understanding the consistency of host intestinal microbial pathophysiological mechanisms in various diseases. Also, our understanding of the impact of gut flora on human health and disease is therefore expected to be widening and deepening.

What merits attention is the distribution of Core-DREEM gene numbers among various diseases (see Figure 5.3). Large amounts of the unique DREEM genes in CD or UC suggest that these two types of IBDs are quite different diseases. However, they share the most abundant markers (118,282 genes) compared with other pairs of diseases, indicating the strong correlation between these two types of IBDs. Moreover, the fact that CD shares more genes with liver cirrhosis than UC does is consistent with the fact that CD can affect any part of the gastrointestinal tract, while UC is restricted to the mucosa[199], as the alterations in composition of gut microbiota, mucosal and systemic immunity, and increase in small intestine bacterial and permeability of small bowel are all potential mechanisms of gut-liver interaction[200]. Considering the awkward situation that the accurate diagnosis of CD and UC is still a challenge[201], these DREEM genes would be valuable for establishing reliable biomarkers for the early and better diagnosis and prognosis of various IBDs. Furthermore, those shared marker genes between obesity and other diseases may help resolve the mechanism how obesity influences other relevant diseases.

Annotation of the DREEM genes provides a picture for the universally microbial taxonomic distribution and metabolic genes involved in the six types of diseases. For instance, the taxonomic classification clearly indicates that the millions of DREEM genes fall in only 581 species, and the most abundant species is *Coprococcus catus*, which has been reported significantly associated with obesity[202]. Another relative abundant species *Clostridium perfringens* has been reported to cause life-threatening gas gangrene and mild enterotoxaemia in human[203]. At phylum level, Proteobacteria are the largest subgroup, which take part 43.0% of all the DREEM genes. This finding emphasizes the medical importance of various members of Proteobacteria in human gastrointestinal disease, which has been well documented[204]. Meanwhile, functional annotation of the Core-DREEM genes find-

s that 104 (2.0%) out of 5,100 tend to gather in type IV systems (COG3451, COG3505 and COG3843 with 5,665 COGs as reference), paralleling previous studies which suggested that the protein substrates of type IV systems are important for the virulence of bacterial pathogens[205]. This result indicates that the overexpression of genes involved in this pathway is significantly correlated with multiple diseases. However, only 1,241,386 (63.6%) out of 1,953,046 DREEM genes are annotated with certain functions. The other 36.4% DREEM genes remains functionally unknown, which supplies a reliable reference for further experiment design to replenish the database of disease-related bacterial pathogens.

As the first released database focusing on the disease-related genes of human gut microbiome, DREEM provides wide and deep vision into the microbial genetic diversity related to relevant human diseases. We hope that DREEM could serve as a reference catalogue for future studies of pathophysiological role of gut microbiomes in human health. Moreover, we expect that, based on intestinal microbiota and with the help of the DREEM database, the new diagnostic and therapeutic strategies of host disease could be invented. The Core-DREEM genes have shown potential usefulness of designing microbiota-targeted biomarkers, which may be a powerful tool for disease detection and treatment.

Nevertheless, the increasing expansion of related projects will be expected to improve our capacity in the future, to compile the latest sequenced samples of more relevant human diseases, such as depression, IBS and so on. We will keep updating so that the database stays current as new disease-related gut metagenomic study published. We plan to evolve DREEM by adding significant functionality. One valuable direction of ongoing development is to establish disease prediction models to help identify relevant diseases based on the DREEM genes. As DREEM supplies references for designing animal experiment model to explore whether these effects interact in ways that influence outcome, drug targets may also be recommended over time. To sum up, we believe that the resource built by DREEM will expand to well meet the further requirements of the research community for human gut microbiomes.

CHAPTER 6
CONCLUSIONS AND FUTURE WORKS

6.1 Conclusions

Focusing on human gut microbiota, this research work covered the most advanced aspects about the metagenomic data analysis of human gut microbiota, and explored the possibility of putting the findings about disease-related human microbiota into application.

In Chapter 2 of this dissertation, we pre-processed and carried out a uniform annotation of the raw data of human gut microbiota from hosts suffering various diseases by applying the state-of-the-art bioinformatics tools. For systematic analysis, a novel binning unit was defined, functional taxonomic unit. With the annotation result, we answered the question of how the ecological niches of gut microbiota correlate with the host health in every step of a well-designed meta-analysis, covering all the four aspects , i.e. taxonomic composition, functional carriage of these microbes, taxonomic co-occurrence network and also functional gene-gene interaction network. Universal taxonomic and functional biomarkers were identified. Interesting finding from the gene-gene interaction network and significantly alteration of taxonomic network patterns indicated that the gut microbiota inside human gut aggregate a 'super organism' and influence the host health in a community manner. In summary, taxonomic composition, microbial functions, microbial correlations and the interactions of microbial functions are four indispensable components for characterizing microbial community, which should be the comprehensive way for defining a pan-microbiome. This literal definition of pan-microbiome provides a practical framework for designing future research works.

In Chapter 3 of this dissertation, we proved the existence of an aging progression of human gut microbiota by applying unsupervised machine learning approaches on metage-

nomics data. We applied an unsupervised machine learning approach SPD on genera abundance profile of human gut microbiota quantified by 16S rRNA sequencing data. Without using the age information of the samples, SPD sorted sample groups on a minimal spanning tree that recapitulated the aging progression. This result indicated the existence of an aging progression reflected in the human gut microbiota. In the meantime, we found 35 genera associated with this age-related progression. Some of these genera were not identified using the commonly-used statistical approaches for metagenomics analysis. Literature review of these 35 genera led to a lot of evidences of the functional relevance of these genera. The evidences collectively indicated an age-related decline of the beneficial functions of gut microbiota, as well as increase of inflammation and diseases, especially for the elderly people older than 90s. This is the first study characterizing the human gut microbiota in a trajectory manner, which sheds light on the possibility of exploring diverse approaches for conducting metagenomics analysis.

In Chapter 4 of this dissertation, we further explored to develop a machine-learning based tool LightCUD in Chapter 4, which was designed to assist the diagnosis of IBD based on human gut microbiome. The well-designed feature selection steps and comparison of different machine learning algorithms contributed to a high-performance tool. Regarding the high-speed development and popularity of NGS, LightCUD highlights the potential of diagnostic tools developed with machine learning algorithms based on the data of human gut microbiome.

In Chapter 5 of this dissertation, we released the first database integrating disease-related genes of human gut microbiota, named DREEM, which provides a clue and data resources for those studies about disease-related changes of gut microbiota.

6.2 Future plan

6.2.1 Interplay between host genome and microbiome

It is clear that the gut microbiota and the host form a closely cross-talked symbiot. It has been proved that both of the taxonomic composition and functional carriage of gut microbiota are correlated with the host metabolites[206]. With 16S rDNA sequencing data of the gut microbiota, it will be possible to characterize the taxonomic composition and community assembly of the microbiota. With whole genome sequencing data, additional information about functional carriage of the microbiota will be available, also finer annotation of the taxonomic information. From the host side, with metabolomics, the faecal or blood metabolites could be quantified. With proteomics, it is possible to quantify the expression levels of thousands of proteins for the certain tissue part. With genomics (bulk or single cell RNA sequencing), the expression level of tens of thousands of genes could be revealed. The marriage of the host metabolites/proteome/gene expression and the fully characterization of the gut microbiota will definitely contribute to series of interesting findings, which will advance our understanding about how the gut microbiota are correlated with the host metabolism.

6.2.2 Detection of microbial association networks in human gut microbiota

As pointed out in Chapter 2, the community structure matters more than individual microbes when characterizing different gut microbial ecosystem. By far, most of the published works tried to reveal microbial assembly rules using cross-sample variation. The variation of different host conditions could confound the findings and lead to false conclusions. So, a well-designed long-term monitoring of host and collecting stool samples at different time points will definitely contribute to a better set of samples for doing correlation analysis to reveal the real community assembly rules of the gut microbiota.Also the methods for revealing the correlation between different microbial compositions is lim-

ited. Exploring new methods for computing the correlation and discovering solid microbial communities is also in an urgent call.

6.2.3 Human gut microbiota aging clocks based on machine learning algorithms

We have tried to develop the machine learning tools LightCUD (Chapter 4) based on human gut microbiota for assisting disease diagnosis. Also, we have proved the existence of an aging progression of human gut microbiota (Chapter 3). The next plan for continuing this dissertation is to develop a human gut microbiota aging clocks based on machine learning algorithms, which could predict the ages of one's gut microbiota and also assisting the designing of microbiome-targeted therapeutic strategies to prevent aging.

REFERENCES

[1] A. v. Leeuwenhoek, "An abstract of a letter from antonie van leeuwenhoek, sep. 12, 1683.," *Philosophical Transactions of the Royal Society*, vol. 14, pp. 568–574,

[2] C. Dobell, "The discovery of the intestinal protozoa of man.," *Proceedings of the Royal Society of Medicine*, vol. 13, no. Sect_Hist_Med, pp. 1–15, 1920.

[3] J. Leidy, *A flora and fauna within living animals*. Smithsonian Institution, 1853, vol. 44.

[4] L Pasteur, "Observations relatives à la note précédente de m," *Duclaux. CR Acad. Sci*, vol. 100, p. 68, 1885.

[5] J. J. Walsh, *Louis Pasteur*. Catholic Encyclopedia, 1913, vol. 11.

[6] E Metchnikoff, "Sur la lutte des cellules de l'organisme contre l'invasion des microbes," *Ann. Inst. Pasteur*, vol. 1, p. 321, 1887.

[7] T. Escherich, *Die darmbakterien des säuglings und ihre beziehungen zur physiologie der Verdauung*. F. Enke, 1886.

[8] A. Nissle, "Über die grundlagen einer neuen ursächlichen bekämpfung der pathologischen darmflora," *Dtsch Med Wochenschr*, vol. 42, pp. 1181–1184, 1916.

[9] S. L. Prescott, "History of medicine: Origin of the term microbiome and why it matters," *Human Microbiome Journal*, vol. 4, pp. 24–25, 2017.

[10] W Lane-Petter, *The provision and use of pathogen-free laboratory animals*, 1962.

[11] J. Whipps, K Lewis, and R. Cooke, "Mycoparasitism and plant disease control," *Fungi in Biological Control Systems*, pp. 161–187, 1988.

[12] J. Lederberg and A. T. McCray, "Ome sweetomics–a genealogical treasury of words," *The Scientist*, vol. 15, no. 7, pp. 8–8, 2001.

[13] R. Hungate, "Chapter iv a roll tube method for cultivation of strict anaerobes," in *Methods in Microbiology*, vol. 3, Elsevier, 1969, pp. 117–132.

[14] B Wostmann and E Bruckner-Kardoss, "Cecal enlargment in germ-free animals," *Am. J. Physiol.*, vol. 197, 1960.

[15] J. T. Staley and A. Konopka, "Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats," *Annual Review of Microbiology*, vol. 39, no. 1, pp. 321–346, 1985.

[16] E. A. Grice and J. A. Segre, "The human microbiome: Our second genome," *Annual Review of Genomics and Human genetics*, vol. 13, pp. 151–170, 2012.

[17] C. Human Microbiome Project, "Structure, function and diversity of the healthy human microbiome," *Nature*, vol. 486, no. 7402, pp. 207–214, 2012.

[18] A. M. O'Hara and F. Shanahan, "The gut flora as a forgotten organ," *EMBO Reports*, vol. 7, no. 7, pp. 688–693, 2006.

[19] F. Bäckhed, R. E. Ley, J. L. Sonnenburg, D. A. Peterson, and J. I. Gordon, "Host-bacterial mutualism in the human intestine," *Science*, vol. 307, no. 5717, pp. 1915–1920, 2005.

[20] F. Guarner and J.-R. Malagelada, "Gut flora in health and disease," *The Lancet*, vol. 361, no. 9356, pp. 512–519, 2003.

[21] R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J.-F. Tomb, B. A. Dougherty, J. M. Merrick, *et al.*, "Whole-genome random sequencing and assembly of haemophilus influenzae rd," *Science*, vol. 269, no. 5223, pp. 496–512, 1995.

[22] E. Kellenberger, "Exploring the unknown," *EMBO reports*, vol. 2, no. 1, pp. 5–7, 2001.

[23] N. R. Pace, "A molecular view of microbial diversity and the biosphere," *Science*, vol. 276, no. 5313, pp. 734–740, 1997.

[24] R. I. Amann, W. Ludwig, and K.-H. Schleifer, "Phylogenetic identification and in situ detection of individual microbial cells without cultivation.," *Microbiol. Mol. Biol. Rev.*, vol. 59, no. 1, pp. 143–169, 1995.

[25] J. C. Wooley, A. Godzik, and I. Friedberg, "A primer on metagenomics," *PLoS Comput Biol*, vol. 6, no. 2, e1000667, 2010.

[26] D. R. Garza and B. E. Dutilh, "From cultured to uncultured genome sequences: Metagenomics and modeling microbial ecosystems," *Cellular and Molecular Life Sciences*, vol. 72, no. 22, pp. 4287–4308, 2015.

[27] J. Aitchison, "The statistical analysis of compositional data," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 44, no. 2, pp. 139–160, 1982.

[28] J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barcelo-Vidal, "Isometric logratio transformations for compositional data analysis," *Mathematical Geology*, vol. 35, no. 3, pp. 279–300, 2003.

[29] M. Greenacre and R. Primicerio, *Multivariate analysis of ecological data*. Fundacion BBVA, 2014.

[30] M. J. Anderson, "A new method for non-parametric multivariate analysis of variance," *Austral ecology*, vol. 26, no. 1, pp. 32–46, 2001.

[31] S. Weiss, W. Van Treuren, C. Lozupone, K. Faust, J. Friedman, Y. Deng, L. C. Xia, Z. Z. Xu, L. Ursell, E. J. Alm, *et al.*, "Correlation detection strategies in microbial data sets vary widely in sensitivity and precision," *The ISME Journal*, vol. 10, no. 7, pp. 1669–1681, 2016.

[32] F. H. Karlsson, V. Tremaroli, I. Nookaew, G. Bergstrom, C. J. Behre, B. Fagerberg, J. Nielsen, and F. Backhed, "Gut metagenome in european women with normal, impaired and diabetic glucose control," *Nature*, vol. 498, no. 7452, pp. 99–103, 2013.

[33] J. Qin, Y. Li, Z. Cai, S. Li, J. Zhu, F. Zhang, S. Liang, W. Zhang, Y. Guan, D. Shen, *et al.*, "A metagenome-wide association study of gut microbiota in type 2 diabetes," *Nature*, vol. 490, no. 7418, pp. 55–60, 2012.

[34] E. A. Franzosa, A. Sirota-Madi, J. Avila-Pacheco, N. Fornelos, H. J. Haiser, S. Reinker, T. Vatanen, A. B. Hall, H. Mallick, L. J. McIver, *et al.*, "Gut microbiome structure and metabolic activity in inflammatory bowel disease," *Nature Microbiology*, vol. 4, no. 2, p. 293, 2019.

[35] E. Le Chatelier, T. Nielsen, J. Qin, E. Prifti, F. Hildebrand, G. Falony, M. Almeida, M. Arumugam, J. M. Batto, S. Kennedy, P. Leonard, J. Li, K. Burgdorf, N. Grarup, T. Jorgensen, I. Brandslund, H. B. Nielsen, A. S. Juncker, M. Bertalan, F. Levenez, N. Pons, S. Rasmussen, S. Sunagawa, J. Tap, S. Tims, E. G. Zoetendal, S. Brunak, K. Clement, J. Dore, M. Kleerebezem, K. Kristiansen, P. Renault, T. Sicheritz-Ponten, W. M. de Vos, J. D. Zucker, J. Raes, T. Hansen, H. I. T. c. Meta, P. Bork, J. Wang, S. D. Ehrlich, and O. Pedersen, "Richness of human gut microbiome correlates with metabolic markers," *Nature*, vol. 500, no. 7464, pp. 541–546, 2013.

[36] Y. Liu, L. Zhang, X. Wang, Z. Wang, J. Zhang, R. Jiang, X. Wang, K. Wang, Z. Liu, Z. Xia, *et al.*, "Similar fecal microbiota signatures in patients with diarrhea-predominant irritable bowel syndrome and patients with depression," *Clin. Gastroenterol. H.*, vol. 14, no. 11, pp. 1602–1611, 2016.

[37] T. Wang, G. Cai, Y. Qiu, N. Fei, M. Zhang, X. Pang, W. Jia, S. Cai, and L. Zhao, "Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers," *The ISME Journal*, vol. 6, no. 2, p. 320, 2012.

[38] O. C. Aroniadis and L. J. Brandt, "Fecal microbiota transplantation: Past, present and future," *Current Opinion in Gastroenterology*, vol. 29, no. 1, pp. 79–84, 2013.

[39] A. Lyra, T. Rinttilä, J. Nikkilä, L. Krogius-Kurikka, K. Kajander, E. Malinen, J. Mättö, L. Mäkelä, and A. Palva, "Diarrhoea-predominant irritable bowel syndrome distinguishable by 16s rrna gene phylotype quantification," *World Journal of Gastroenterology: WJG*, vol. 15, no. 47, p. 5936, 2009.

[40] R. E. Ley, P. J. Turnbaugh, S. Klein, and J. I. Gordon, "Microbial ecology: Human gut microbes associated with obesity," *Nature*, vol. 444, no. 7122, p. 1022, 2006.

[41] M. H. Van Nuenen, K. Venema, J. C. Van Der Woude, and E. J. Kuipers, "The metabolic activity of fecal microbiota from healthy individuals and patients with inflammatory bowel disease," *Digestive Diseases and Sciences*, vol. 49, no. 3, pp. 485–491, 2004.

[42] A. Naseribafrouei, K. Hestad, E. Avershina, M. Sekelja, A. Linløkken, R. Wilson, and K. Rudi, "Correlation between the human fecal microbiota and depression," *Neurogastroenterology & Motility*, vol. 26, no. 8, pp. 1155–1162, 2014.

[43] N. Qin, F. Yang, A. Li, E. Prifti, Y. Chen, L. Shao, J. Guo, E. Le Chatelier, J. Yao, L. Wu, J. Zhou, S. Ni, L. Liu, N. Pons, J. M. Batto, S. P. Kennedy, P. Leonard, C. Yuan, W. Ding, Y. Chen, X. Hu, B. Zheng, G. Qian, W. Xu, S. D. Ehrlich, S. Zheng, and L. Li, "Alterations of the human gut microbiome in liver cirrhosis," *Nature*, vol. 513, no. 7516, pp. 59–64, 2014.

[44] D. A. Relman, "The human microbiome: Ecosystem resilience and health," *Nutrition Reviews*, vol. 70, no. suppl_1, S2–S9, 2012.

[45] L. B. Lovat, "Age related changes in gut physiology and nutritional status," *Gut*, 1996.

[46] G. T. M. M J Hopkins R Sharp, "Age and disease related changes in intestinal bacterial populations assessed by cell culture, 16s rrna abundance, and community cellular fatty acid profiles," *Gut*, 2001.

[47] T. Odamaki, K. Kato, H. Sugahara, N. Hashikura, S. Takahashi, J. Z. Xiao, F. Abe, and R. Osawa, "Age-related changes in gut microbiota composition from newborn to centenarian: A cross-sectional study," *BMC Microbiol*, vol. 16, p. 90, 2016.

[48] T. Yatsunenko, F. E. Rey, M. J. Manary, I. Trehan, M. G. Dominguez-Bello, M. Contreras, M. Magris, G. Hidalgo, R. N. Baldassano, A. P. Anokhin, A. C. Heath, B. Warner, J. Reeder, J. Kuczynski, J. G. Caporaso, C. A. Lozupone, C. Lauber, J. C. Clemente, D. Knights, R. Knight, and J. I. Gordon, "Human gut microbiome viewed across age and geography," *Nature*, vol. 486, no. 7402, pp. 222–7, 2012.

[49] G. Bian, "The gut microbiota of healthy aged chinese is similar to that of the healthy young," *MSphere*, 2017.

[50] C. J. Stewart, N. J. Ajami, J. L. O'Brien, D. S. Hutchinson, D. P. Smith, M. C. Wong, M. C. Ross, R. E. Lloyd, H. Doddapaneni, G. A. Metcalf, D. Muzny, R. A. Gibbs, T. Vatanen, C. Huttenhower, R. J. Xavier, M. Rewers, W. Hagopian, J. Toppari, A. G. Ziegler, J. X. She, B. Akolkar, A. Lernmark, H. Hyoty, K. Vehik, J. P. Krischer, and J. F. Petrosino, "Temporal development of the gut microbiome in early childhood from the teddy study," *Nature*, vol. 562, no. 7728, pp. 583–588, 2018.

[51] Y. Lan, A. Kriete, and G. L. Rosen, "Selecting age-related functional characteristics in the human gut microbiome," *Microbiome*, vol. 1, no. 1, p. 2, 2013.

[52] T. W. Buford, "(dis)trust your gut: The gut microbiome in age-related inflammation, health, and disease," *Microbiome*, vol. 5, no. 1, p. 80, 2017.

[53] E. B. Minelli, A. Benini, A. M. Beghini, R. Cerutti, and G. Nardo, "Bacterial faecal flora in healthy women of different ages," *Microbial Ecology in Health and Disease*, vol. 6, no. 2, pp. 43–51, 1993.

[54] T. Vatanen, E. A. Franzosa, R. Schwager, S. Tripathi, T. D. Arthur, K. Vehik, A. Lernmark, W. A. Hagopian, M. J. Rewers, J. X. She, J. Toppari, A. G. Ziegler, B. Akolkar, J. P. Krischer, C. J. Stewart, N. J. Ajami, J. F. Petrosino, D. Gevers, H. Lahdesmaki, H. Vlamakis, C. Huttenhower, and R. J. Xavier, "The human gut microbiome in early-onset type 1 diabetes from the teddy study," *Nature*, vol. 562, no. 7728, pp. 589–594, 2018.

[55] B. P. Willing, S. L. Russell, and B. B. Finlay, "Shifting the balance: Antibiotic effects on host–microbiota mutualism," *Nature Reviews Microbiology*, vol. 9, no. 4, pp. 233–243, 2011.

[56] A. C. Brown and A. Valiere, "Probiotics and medical nutrition therapy," *Nutrition in Clinical Care: An Official Publication of Tufts University*, vol. 7, no. 2, p. 56, 2004.

[57] S. Bengmark, "Colonic food: Pre-and probiotics," *The American Journal of Gastroenterology*, vol. 95, no. 1, S5–S7, 2000.

[58] A. Evrensel and M. E. Ceylan, "Fecal microbiota transplantation and its usage in neuropsychiatric disorders," *Clinical Psychopharmacology and Neuroscience*, vol. 14, no. 3, p. 231, 2016.

[59] P. Sajda, "Machine learning for detection and diagnosis of disease," *Annu. Rev. Biomed. Eng.*, vol. 8, pp. 537–565, 2006.

[60] D. Knights, E. K. Costello, and R. Knight, "Supervised classification of human microbiota," *FEMS Microbiology Reviews*, vol. 35, no. 2, pp. 343–359, 2011.

[61] C. Duvallet, S. M. Gibbons, T. Gurry, R. A. Irizarry, and E. J. Alm, "Meta-analysis of gut microbiome studies identifies disease-specific and shared responses," *Nature Communications*, vol. 8, no. 1, pp. 1–10, 2017.

[62] E. Pasolli, D. T. Truong, F. Malik, L. Waldron, and N. Segata, "Machine learning meta-analysis of large metagenomic datasets: Tools and biological insights," *PLoS Computational Biology*, vol. 12, no. 7, 2016.

[63] F. Guarner and J.-R. Malagelada, "Gut flora in health and disease," *Lancet*, vol. 361, no. 9356, pp. 512–519, 2003.

[64] N. W. Griffin, P. P. Ahern, J. Cheng, A. C. Heath, O. Ilkayeva, C. B. Newgard, L. Fontana, and J. I. Gordon, "Prior dietary practices and connections to a human gut microbial metacommunity alter responses to diet interventions," *Cell Host & Microbe*, vol. 21, no. 1, pp. 84–96, 2017.

[65] Y. Lan, A. Kriete, and G. L. Rosen, "Selecting age-related functional characteristics in the human gut microbiome," *Microbiome*, vol. 1, no. 1, p. 2, 2013.

[66] J. Li, H. Jia, X. Cai, H. Zhong, Q. Feng, S. Sunagawa, M. Arumugam, J. R. Kultima, E. Prifti, T. Nielsen, *et al.*, "An integrated catalog of reference genes in the human gut microbiome," *Nat. Biotechnol.*, vol. 32, no. 8, pp. 834–841, 2014.

[67] J. Qin, R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, D. R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J.-M. Batto, T. Hansen, D. Le Paslier, A. Linneberg, H. B. Nielsen, E. Pelletier, P. Renault, T. Sicheritz-Ponten, K. Turner, H. Zhu, C. Yu, S. Li, M. Jian, Y. Zhou, Y. Li, X. Zhang, S. Li, N. Qin, H. Yang, J. Wang, S. Brunak, J. Dor, F. Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, M. Antolin, F. Artiguenave, H. Blottiere, N. Borruel, T. Bruls, F. Casellas, C. Chervaux, A. Cultrone, C. Delorme, G. Denariaz, R. Dervyn, M. Forte, C. Friss, M. van de Guchte, E. Guedon, F. Haimet, A. Jamet, C. Juste, G. Kaci, M. Kleerebezem, J. Knol, M. Kristensen, S. Layec, K. Le Roux, M. Leclerc, E. Maguin, R. Melo Minardi, R. Oozeer, M. Rescigno, N. Sanchez, S. Tims, T. Torrejon, E. Varela, W. de Vos, Y. Winogradsky, E. Zoetendal, P. Bork,

S. D. Ehrlich, and J. Wang, "A human gut microbial gene catalogue established by metagenomic sequencing," *Nature*, vol. 464, no. 7285, pp. 59–65, 2010.

[68] M. Deschasaux, K. E. Bouter, A. Prodan, E. Levin, A. K. Groen, H. Herrema, V. Tremaroli, G. J. Bakker, I. Attaye, S.-J. Pinto-Sietsma, *et al.*, "Depicting the composition of gut microbiota in a population with varied ethnic origins but shared geography," *Nature Medicine*, vol. 24, no. 10, p. 1526, 2018.

[69] Y. He, W. Wu, H.-M. Zheng, P. Li, D. McDonald, H.-F. Sheng, M.-X. Chen, Z.-H. Chen, G.-Y. Ji, P. Mujagond, *et al.*, "Regional variation limits applications of healthy gut microbiome reference ranges and disease models," *Nature Medicine*, vol. 24, no. 10, p. 1532, 2018.

[70] C. Duvallet, S. M. Gibbons, T. Gurry, R. A. Irizarry, and E. J. Alm, "Meta-analysis of gut microbiome studies identifies disease-specific and shared responses," *Nature Communications*, vol. 8, no. 1, p. 1784, 2017.

[71] M. A. Sze and P. D. Schloss, "Looking for a signal in the noise: Revisiting obesity and the microbiome," *MBio*, vol. 7, no. 4, 2016.

[72] W. A. Walters, Z. Xu, and R. Knight, "Meta-analyses of human gut microbes associated with obesity and ibd," *FEBS Letters*, vol. 588, no. 22, pp. 4223–4233, 2014.

[73] X. Jiang, X. Li, L. Yang, C. Liu, Q. Wang, W. Chi, and H. Zhu, "How microbes shape their communities? a microbial community model based on functional genes," *Genomics, proteomics & bioinformatics*, vol. 17, no. 1, pp. 91–105, 2019.

[74] J. Qin, R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, *et al.*, "A human gut microbial gene catalogue established by metagenomic sequencing," *Nature*, vol. 464, no. 7285, p. 59, 2010.

[75] M. Deschasaux, K. E. Bouter, A. Prodan, E. Levin, A. K. Groen, H. Herrema, V. Tremaroli, G. J. Bakker, I. Attaye, S. J. Pinto-Sietsma, D. H. van Raalte, M. B. Snijder, M. Nicolaou, R. Peters, A. H. Zwinderman, F. Backhed, and M. Nieuwdorp, "Depicting the composition of gut microbiota in a population with varied ethnic origins but shared geography," *Nat Med*, vol. 24, no. 10, pp. 1526–1531, 2018.

[76] Y. He, W. Wu, H. M. Zheng, P. Li, D. McDonald, H. F. Sheng, M. X. Chen, Z. H. Chen, G. Y. Ji, Z. D. Zheng, P. Mujagond, X. J. Chen, Z. H. Rong, P. Chen, L. Y. Lyu, X. Wang, C. B. Wu, N. Yu, Y. J. Xu, J. Yin, J. Raes, R. Knight, W. J. Ma, and H. W. Zhou, "Regional variation limits applications of healthy gut microbiome reference ranges and disease models," *Nat Med*, vol. 24, no. 10, pp. 1532–1535, 2018.

[77] M. M. Finucane, T. J. Sharpton, T. J. Laurent, and K. S. Pollard, "A taxonomic signature of obesity in the microbiome? getting to the guts of the matter," *PloS One*, vol. 9, no. 1, e84689, 2014.

[78] H. B. Nielsen, M. Almeida, A. S. Juncker, S. Rasmussen, J. Li, S. Sunagawa, D. R. Plichta, L. Gautier, A. G. Pedersen, E. Le Chatelier, E. Pelletier, I. Bonde, T. Nielsen, C. Manichanh, M. Arumugam, J. M. Batto, M. B. Quintanilha Dos Santos, N. Blom, N. Borruel, K. S. Burgdorf, F. Boumezbeur, F. Casellas, J. Dore, P. Dworzynski, F. Guarner, T. Hansen, F. Hildebrand, R. S. Kaas, S. Kennedy, K. Kristiansen, J. R. Kultima, P. Leonard, F. Levenez, O. Lund, B. Moumen, D. Le Paslier, N. Pons, O. Pedersen, E. Prifti, J. Qin, J. Raes, S. Sorensen, J. Tap, S. Tims, D. W. Ussery, T. Yamada, H. I. T. C. Meta, P. Renault, T. Sicheritz-Ponten, P. Bork, J. Wang, S. Brunak, S. D. Ehrlich, and H. I. T. C. Meta, "Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes," *Nat. Biotechnol.*, vol. 32, no. 8, pp. 822–828, 2014.

[79] J. D. Lewis, E. Z. Chen, R. N. Baldassano, A. R. Otley, A. M. Griffiths, D. Lee, K. Bittinger, A. Bailey, E. S. Friedman, C. Hoffmann, L. Albenberg, R. Sinha, C. Compher, E. Gilroy, L. Nessel, A. Grant, C. Chehoud, H. Li, G. D. Wu, and F. D. Bushman, "Inflammation, antibiotics, and diet as environmental stressors of the gut microbiome in pediatric crohn's disease," *Cell Host Microbe*, vol. 18, no. 4, pp. 489–500, 2015.

[80] F. H. Karlsson, F. Fak, I. Nookaew, V. Tremaroli, B. Fagerberg, D. Petranovic, F. Backhed, and J. Nielsen, "Symptomatic atherosclerosis is associated with an altered gut metagenome," *Nat. Commun.*, vol. 3, p. 1245, 2012.

[81] S. K. P. Peng Qiu Andrew J. Gentles, "Discovering biological progression underlying microarray samples," *PLoS Comput Biol*, 2011.

[82] P. Qiu, E. F. Simonds, S. C. Bendall, J. Gibbs K. D., R. V. Bruggner, M. D. Linderman, K. Sachs, G. P. Nolan, and S. K. Plevritis, "Extracting a cellular hierarchy from high-dimensional cytometry data with spade," *Nat Biotechnol*, vol. 29, no. 10, pp. 886–91, 2011.

[83] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn, "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells," *Nature Biotechnology*, vol. 32, no. 4, p. 381, 2014.

[84] J. L, R. S, W. RK, D. RH, and M. DP, "Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease," *Nature*, vol. 491, pp. 119–124, 2012.

[85] J. M. Dahlhamer, E. P. Zammitti, B. W. Ward, A. G. Wheaton, and J. B. Croft, "Prevalence of inflammatory bowel disease among adults aged 18 years united states, 2015," *MMWR. Morbidity and Mortality Weekly Report*, vol. 65, no. 42, pp. 1166–1169, 2016.

[86] Q. Ouyang, R. Tandon, K. L. Goh, G.-Z. Pan, K. M. Fock, C. Fiocchi, S. K. Lam, and S.-D. Xiao, "Management consensus of inflammatory bowel disease for the asia pacific region," *Journal of Gastroenterology and Hepatology*, vol. 21, no. 12, pp. 1772–1782, 2006.

[87] D. C. Baumgart, "The diagnosis and treatment of crohn's disease and ulcerative colitis," *Deutsches Aerzteblatt Int*, 2009.

[88] S. Kugathasan and C. Fiocchi, "Progress in basic inflammatory bowel disease research," *Seminars in Pediatric Surgery*, vol. 16, no. 3, pp. 146–153, 2007.

[89] J. D. Lewis, "The utility of biomarkers in the diagnosis and therapy of inflammatory bowel disease," *Gastroenterology*, vol. 140, no. 6, 1817–1826.e2, 2011.

[90] J. Tibble, "A simple method for assessing intestinal inflammation in crohn's disease," *Gut*, vol. 47, no. 4, pp. 506–513, 2000.

[91] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, 2017, pp. 3146–3154.

[92] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014, ISBN: 1139952749.

[93] N. M. Nasrabadi, "Pattern recognition and machine learning," *Journal of electronic imaging*, vol. 16, no. 4, p. 049 901, 2007.

[94] Z. John Lu, "The elements of statistical learning: Data mining, inference, and prediction," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 173, no. 3, pp. 693–694, 2010.

[95] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, "Genbank," *Nucleic Acids Research*, vol. 41, no. D1, pp. D36–D42, 2012.

[96] G. Stoesser, "The embl nucleotide sequence database," *Nucleic Acids Research*, vol. 30, no. 1, pp. 21–26, 2002.

[97] A. M. O'Hara and F. Shanahan, "The gut flora as a forgotten organ," *EMBO Rep.*, vol. 7, no. 7, pp. 688–693, 2006.

[98] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, "Genbank," *Nucleic Acids Research*, vol. 41, no. D1, pp. D36–D42, 2013.

[99] G. Stoesser, W. Baker, A. van den Broek, E. Camon, M. Garcia-Pastor, C. Kanz, T. Kulikova, R. Leinonen, Q. Lin, V. Lombard, *et al.*, "The embl nucleotide sequence database," *Nucleic Acids Research*, vol. 30, no. 1, pp. 21–26, 2002.

[100] B. Lai, F. Wang, X. Wang, L. Duan, and H. Zhu, "Intemap: Integrated metagenomic assembly pipeline for ngs short reads," *BMC Bioinformatics*, vol. 16, no. 1, p. 244, 2015.

[101] W. Zhu, A. Lomsadze, and M. Borodovsky, "Ab initio gene identification in metagenomic sequences," *Nucleic Acids Research*, vol. 38, no. 12, e132–e132, 2010.

[102] Y. Liu, J. Guo, G. Hu, and H. Zhu, "Gene prediction in metagenomic fragments based on the svm algorithm," *BMC Bioinformatics*, vol. 14, no. 5, S12, 2013.

[103] G.-Q. Hu, J.-T. Guo, Y.-C. Liu, and H. Zhu, "Metatisa: Metagenomic translation initiation site annotator for improving gene start prediction," *Bioinformatics*, vol. 25, no. 14, pp. 1843–1845, 2009.

[104] W. Li and A. Godzik, "Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, 2006.

[105] A. Brady and S. L. Salzberg, "Phymm and phymmbl: Metagenomic phylogenetic classification with interpolated markov models," *Nat. Methods*, vol. 6, no. 9, pp. 673–676, 2009.

[106] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped blast and psi-blast: A new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[107] R. L. Tatusov, M. Y. Galperin, D. A. Natale, and E. V. Koonin, "The cog database: A tool for genome-scale analysis of protein functions and evolution," *Nucleic Acids Research*, vol. 28, no. 1, pp. 33–36, 2000.

[108] M. Kanehisa and S. Goto, "Kegg: Kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000.

[109] K. D. Pruitt, T. Tatusova, and D. R. Maglott, "Ncbi reference sequences (refseq): A curated non-redundant sequence database of genomes, transcripts and proteins," *Nucleic Acids Research*, vol. 35, no. suppl_1, pp. D61–D65, 2006.

[110] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with bowtie 2," *Nat. Mmethods*, vol. 9, no. 4, pp. 357–359, 2012.

[111] F. Wilcoxon, S. Katti, and R. A. Wilcox, "Critical values and probability levels for the wilcoxon rank sum test and the wilcoxon signed rank test," *Sel. Tbl. Math. STA.*, vol. 1, pp. 171–259, 1970.

[112] J. Friedman and E. J. Alm, "Inferring correlation networks from genomic survey data," *PLoS computational biology*, vol. 8, no. 9, e1002687, 2012.

[113] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: A software environment for integrated models of biomolecular interaction networks," *Genome Res.*, vol. 13, no. 11, pp. 2498–2504, 2003.

[114] D. Bonchev and G. A. Buck, "Quantitative measures of network complexity," in *Complexity in chemistry, biology, and ecology*, Springer, 2005, pp. 191–235.

[115] T. Yatsunenko, F. E. Rey, M. J. Manary, I. Trehan, M. G. Dominguez-Bello, M. Contreras, M. Magris, G. Hidalgo, R. N. Baldassano, A. P. Anokhin, *et al.*, "Human gut microbiome viewed across age and geography," *Nature*, vol. 486, no. 7402, p. 222, 2012.

[116] S. B. Barbuddhe and T. Chakraborty, "Listeria as an enteroinvasive gastrointestinal pathogen," in *Molecular mechanisms of bacterial infection via the gut*, Springer, 2009, pp. 173–195.

[117] K. L. Tkaczuk, I. A Shumilin, M. Chruszcz, E. Evdokimova, A. Savchenko, and W. Minor, "Structural and functional insight into the universal stress protein family," *Evol. Appl.*, vol. 6, no. 3, pp. 434–449, 2013.

[118] W.-T. Liu, M. H. Karavolos, D. M. Bulmer, A. Allaoui, R. D. C. E. Hormaeche, J. J. Lee, and C. A. Khan, "Role of the universal stress protein uspa of salmonella in growth arrest, stress and virulence," *Microb. Pathogenesis*, vol. 42, no. 1, pp. 2–10, 2007.

[119] K. J. Ryan and C. G. Ray, "Medical microbiology," *McGraw Hill*, vol. 4, p. 370, 2004.

[120] J. Deutscher, F. M. D. Aké, M. Derkaoui, A. C. Zébré, T. N. Cao, H. Bouraoui, T. Kentache, A. Mokhtari, E. Milohanic, and P. Joyet, "The bacterial phosphoenolpyruvate: Carbohydrate phosphotransferase system: Regulation by protein phosphorylation and phosphorylation-dependent protein-protein interactions," *Microbiol. Mol. Biol. R.*, vol. 78, no. 2, pp. 231–256, 2014.

[121] A. Zelezniak, S. Andrejev, O. Ponomarova, D. R. Mende, P. Bork, and K. R. Patil, "Metabolic dependencies drive species co-occurrence in diverse microbial communities," *Proceedings of the National Academy of Sciences*, vol. 112, no. 20, pp. 6449–6454, 2015.

[122] T. Woyke, H. Teeling, N. N. Ivanova, M. Huntemann, M. Richter, F. O. Gloeckner, D. Boffelli, K. W. Barry, H. J. Shapiro, I. J. Anderson, *et al.*, "Symbiosis insights through metagenomic analysis of a microbialconsortium," *Nature*, vol. 443, no. LBNL–60435, 2006.

[123] R. Levy and E. Borenstein, "Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, no. 31, pp. 12 804–12 809, 2013.

[124] D. R. Plichta, A. S. Juncker, M. Bertalan, E. Rettedal, L. Gautier, E. Varela, C. Manichanh, C. Fouqueray, F. Levenez, T. Nielsen, *et al.*, "Transcriptional interactions suggest niche segregation among microorganisms in the human gut," *Nature Microbiology*, vol. 1, p. 16 152, 2016.

[125] S. Moens and J. Vanderleyden, "Functions of bacterial flagella," *Crit. Rev. Microbiol.*, vol. 22, no. 2, pp. 67–100, 1996.

[126] X. Wang and P. J. Quinn, "Lipopolysaccharide: Biosynthetic pathway and structure modification," *Prog. Lipid. Res.*, vol. 49, no. 2, pp. 97–107, 2010.

[127] A. H. Moeller, S. Foerster, M. L. Wilson, A. E. Pusey, B. H. Hahn, and H. Ochman, "Social behavior shapes the chimpanzee pan-microbiome," *Sci. Adv.*, vol. 2, no. 1, e1500997, 2016.

[128] M. H. Leung and P. K. Lee, "The roles of the outdoors and occupants in contributing to a potential pan-microbiome of the built environment: A review," *Microbiome*, vol. 4, no. 1, p. 21, 2016.

[129] M. H. Leung, D. Wilkins, and P. K. Lee, "Insights into the pan-microbiome: Skin microbial communities of chinese individuals differ from other racial groups," *Sci. Rep.*, vol. 5, p. 11 845, 2015.

[130] V. Bocci, "The neglected organ: Bacterial flora has a crucial immunostimulatory role," *Perspect Biol Med*, vol. 35, no. 2, pp. 251–260, 1992.

[131] N. W. Griffin, P. P. Ahern, J. Cheng, A. C. Heath, O. Ilkayeva, C. B. Newgard, L. Fontana, and J. I. Gordon, "Prior dietary practices and connections to a human gut microbial metacommunity alter responses to diet interventions," *Cell Host Microbe*, vol. 21, no. 1, pp. 84–96, 2017.

[132] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn, "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells," *Nat Biotechnol*, vol. 32, no. 4, p. 381, 2014.

[133] P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, and C. J. Robinson, "Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities," *J Appl Environ Microbiol*, vol. 75, no. 23, pp. 7537–7541, 2009.

[134] C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glckner, "The silva ribosomal rna gene database project: Improved data processing and web-based tools," *Nucleic Acids Research*, vol. 41, no. D1, pp. D590–D596, 2012.

[135] M. J. Anderson, "Permutation tests for univariate or multivariate analysis of variance and regression," *Can J Fish Aquat Sci*, vol. 58, no. 3, pp. 626–639, 2001.

[136] S. H. Duncan, A. J. Richardson, P. Kaul, R. P. Holmes, M. J. Allison, and C. S. Stewart, "Oxalobacter formigenes and its potential role in human health," *J Appl Environ Microbiol*, vol. 68, no. 8, pp. 3841–3847, 2002.

[137] B. Sovran, F. Hugenholtz, M. Elderman, A. A. Van Beek, K. Graversen, M. Huijskes, M. V. Boekschoten, H. F. Savelkoul, P. De Vos, and J. Dekker, "Age-associated impairment of the mucus barrier function is associated with profound changes in microbiota and immunity," *Sci Rep*, vol. 9, no. 1, p. 1437, 2019.

[138] R. Vemuri, T. Shinde, R. Gundamaraju, S. Gondalia, A. Karpe, D. Beale, C. Martoni, and R. Eri, "Lactobacillus acidophilus dds-1 modulates the gut microbiota and improves metabolic profiles in aging mice," *Nutrients*, vol. 10, no. 9, p. 1255, 2018.

[139] N. Tuikhar, S. Keisam, R. K. Labala, P. Ramakrishnan, M. C. Arunkumar, G. Ahmed, E. Biagi, and K. Jeyaram, "Comparative analysis of the gut microbiota in centenarians and young adults shows a common signature across genotypically non-related populations," *Mech Ageing Dev*, 2019.

[140] J. Walter, "Ecological role of lactobacilli in the gastrointestinal tract: Implications for fundamental and biomedical research," *J Appl Environ Microbiol*, vol. 74, no. 16, pp. 4985–4996, 2008.

[141] K. D. Kohl, J. Amaya, C. A. Passement, M. D. Dearing, and M. D. McCue, "Unique and shared responses of the gut microbiota to prolonged fasting: A comparative study across five classes of vertebrate hosts," *FEMS Immunol Med Microbiol*, vol. 90, no. 3, pp. 883–894, 2014.

[142] S. Tims, C. Derom, D. M. Jonkers, R. Vlietinck, W. H. Saris, M. Kleerebezem, W. M. De Vos, and E. G. Zoetendal, "Microbiota conservation and bmi signatures in adult monozygotic twins," *ISME J*, vol. 7, no. 4, p. 707, 2013.

[143] J. K. Goodrich, J. L. Waters, A. C. Poole, J. L. Sutter, O. Koren, R. Blekhman, M. Beaumont, W. Van Treuren, R. Knight, and J. T. Bell, "Human genetics shape the gut microbiome," *Cell*, vol. 159, no. 4, pp. 789–799, 2014.

[144] F. J. Verdam, S. Fuentes, C. de Jonge, E. G. Zoetendal, R. Erbil, J. W. Greve, W. A. Buurman, W. M. de Vos, and S. S. Rensen, "Human intestinal microbiota composition is associated with local and systemic inflammation in obesity," *Obesity (Silver Spring)*, vol. 21, no. 12, E607–15, 2013.

[145] J. S. Escobar, B. Klotz, B. E. Valdes, and G. M. Agudelo, "The gut microbiota of colombians differs from that of americans, europeans and asians," *BMC Microbiol*, vol. 14, p. 311, 2014.

[146] Z. Xu and R. Knight, "Dietary effects on human gut microbiome diversity," *Br J Nutr*, vol. 113, no. S1, S1–S5, 2015.

[147] M. M. Unger, J. Spiegel, K.-U. Dillmann, D. Grundmann, H. Philippeit, J. Brmann, K. Fabender, A. Schwiertz, and K.-H. Schfer, "Short chain fatty acids and gut microbiota differ between patients with parkinson's disease and age-matched controls," *Parkinsonism Relat Disord*, vol. 32, pp. 66–72, 2016.

[148] A. O'Callaghan and D. van Sinderen, "Bifidobacteria and their role as members of the human gut microbiota," *Front Microbial*, vol. 7, p. 925, 2016.

[149] A. Barcenilla, S. E. Pryde, J. C. Martin, S. H. Duncan, C. S. Stewart, C. Henderson, and H. J. Flint, "Phylogenetic relationships of butyrate-producing bacteria from the human gut," *J Appl Environ Microbiol*, vol. 66, no. 4, pp. 1654–1661, 2000.

[150] S. Zou, L. Fang, and M. H. Lee, "Dysbiosis of gut microbiota in promoting the development of colorectal cancer," *Gastroenterol Rep (Oxf)*, vol. 6, no. 1, pp. 1–12, 2018.

[151] B. Flemer, D. B. Lynch, J. M. Brown, I. B. Jeffery, F. J. Ryan, M. J. Claesson, M. O'Riordain, F. Shanahan, and P. W. O'Toole, "Tumour-associated and non-tumour-associated microbiota in colorectal cancer," *Gut*, vol. 66, no. 4, pp. 633–643, 2017.

[152] G. Nakatsu, X. Li, H. Zhou, J. Sheng, S. H. Wong, W. K. Wu, S. C. Ng, H. Tsoi, Y. Dong, N. Zhang, Y. He, Q. Kang, L. Cao, K. Wang, J. Zhang, Q. Liang, J. Yu, and J. J. Sung, "Gut mucosal microbiome across stages of colorectal carcinogenesis," *Nature Communications*, vol. 6, p. 8727, 2015.

[153] W. Chen, F. Liu, Z. Ling, X. Tong, and C. Xiang, "Human intestinal lumen and mucosa-associated microbiota in patients with colorectal cancer," *PloS One*, vol. 7, no. 6, e39743, 2012.

[154] M. S. Zinkernagel, D. C. Zysset-Burri, I. Keller, L. E. Berger, A. B. Leichtle, C. R. Largiader, G. M. Fiedler, and S. Wolf, "Association of the intestinal microbiome with the development of neovascular age-related macular degeneration," *Sci Rep*, vol. 7, p. 40 826, 2017.

[155] F. Strati, D. Cavalieri, D. Albanese, C. De Felice, C. Donati, J. Hayek, O. Jousson, S. Leoncini, D. Renzi, A. Calabro, and C. De Filippo, "New evidences on the altered gut microbiota in autism spectrum disorders," *Microbiome*, vol. 5, no. 1, p. 24, 2017.

[156] R. Padmanabhan, G. Dubourg, J. C. Lagier, C. Couderc, C. Michelle, D. Raoult, and P. E. Fournier, "Genome sequence and description of corynebacterium ihumii sp. nov," *Stand Genomic Sci*, vol. 9, no. 3, pp. 1128–43, 2014.

[157] S. Li, Z. Wang, Y. Yang, S. Yang, C. Yao, K. Liu, S. Cui, Q. Zou, H. Sun, and G. Guo, "Lachnospiraceae shift in the microbial community of mice faecal sample effects on water immersion restraint stress," *AMB Express*, vol. 7, no. 1, p. 82, 2017.

[158] S. Devkota, Y. Wang, M. W. Musch, V. Leone, H. Fehlner-Peach, A. Nadimpalli, D. A. Antonopoulos, B. Jabri, and E. B. Chang, "Dietary-fat-induced taurocholic acid promotes pathobiont expansion and colitis in il10$-/-$ mice," *Nature*, vol. 487, no. 7405, pp. 104–8, 2012.

[159] J. Loubinoux, J.-P. Bronowicki, I. A. Pereira, J.-L. Mougenel, and A. E. Le Faou, "Sulfate-reducing bacteria in human feces and their association with inflammatory bowel diseases," *FEMS Immunol Med Microbiol*, vol. 40, no. 2, pp. 107–112, 2002.

[160] Z. Feng, W. Long, B. Hao, D. Ding, X. Ma, L. Zhao, and X. Pang, "A human stool-derived bilophila wadsworthia strain caused systemic inflammation in specific-pathogen-free mice," *Gut Pathog*, vol. 9, no. 1, p. 59, 2017.

[161] J.-H. Shin, M. Sim, J.-Y. Lee, and D.-M. Shin, "Lifestyle and geographic insights into the distinct gut microbiota in elderly women from two different geographic locations," *J Physiol Anthropol*, vol. 35, no. 1, p. 31, 2016.

[162] V. J. Maffei, S. Kim, E. Blanchard IV, M. Luo, S. M. Jazwinski, C. M. Taylor, and D. A. Welsh, "Biological aging and the human gut microbiota," *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences*, vol. 72, no. 11, pp. 1474–1482, 2017.

[163] J de la Cuesta-Zuluaga, S. Kelley, Y Chen, J. Escobar, N. Mueller, R. Ley, D McDonald, S Huang, A. Swafford, R Knight, *et al.*, *Age-and sex-dependent patterns of gut microbial diversity in human adults. msystems 4: E00261-19*, 2019.

[164] T. Lobatón, I. Hoffman, S. Vermeire, M. Ferrante, J. Verhaegen, and G. Van Assche, "Aeromonas species: An opportunistic enteropathogen in patients with inflammatory bowel diseases? a single center cohort study," *Inflammatory bowel diseases*, vol. 21, no. 1, pp. 71–78, 2014.

[165] U. Gophna, K. Sommerfeld, S. Gophna, W. F. Doolittle, and S. J. V. van Zanten, "Differences between tissue-associated intestinal microfloras of patients with crohn's disease and ulcerative colitis," *Journal of clinical microbiology*, vol. 44, no. 11, pp. 4136–4141, 2006.

[166] S. R. Dalal and E. B. Chang, "The microbial basis of inflammatory bowel diseases," *The Journal of clinical investigation*, vol. 124, no. 10, pp. 4190–4196, 2014.

[167] R. Padmanabhan, G. Dubourg, J.-C. Lagier, C. Couderc, C. Michelle, D. Raoult, and P.-E. Fournier, "Genome sequence and description of corynebacterium ihumii sp. nov.," *Standards in genomic sciences*, vol. 9, no. 3, p. 1128, 2014.

[168] A. Feeney, K. A. Kropp, R. OConnor, and R. D. Sleator, "Cronobacter sakazakii: Stress survival and virulence potential in an opportunistic foodborne pathogen," *Gut microbes*, vol. 5, no. 6, pp. 711–718, 2014.

[169] M. Zeng, N Inohara, and G Nuñez, "Mechanisms of inflammation-driven bacterial dysbiosis in the gut," *Mucosal immunology*, vol. 10, no. 1, p. 18, 2017.

[170] C. M. Guinane and P. D. Cotter, "Role of the gut microbiota in health and chronic gastrointestinal disease: Understanding a hidden metabolic organ," *Therapeutic advances in gastroenterology*, vol. 6, no. 4, pp. 295–308, 2013.

[171] H. Liu, J. Zhu, Q. Hu, and X. Rao, "Morganella morganii, a non-negligent opportunistic pathogen," *International Journal of Infectious Diseases*, vol. 50, pp. 10–17, 2016.

[172] A. L. Hamilton, M. A. Kamm, S. C. Ng, and M. Morrison, "Proteus spp. as putative gastrointestinal pathogens," *Clinical microbiology reviews*, vol. 31, no. 3, e00085–17, 2018.

[173] M. J. Anderson, "Permutation tests for univariate or multivariate analysis of variance and regression," *Canadian journal of fisheries and aquatic scienc es*, vol. 58, no. 3, pp. 626–639, 2001.

[174] J. M. Dahlhamer, E. P. Zammitti, B. W. Ward, A. G. Wheaton, and J. B. Croft, "Prevalence of inflammatory bowel disease among adults aged 18 years united states, 2015," *MMWR. Morbidity and Mortality Weekly Report*, vol. 65, no. 42, pp. 1166–1169, 2016.

[175] Q. Ouyang, R. Tandon, K. L. Goh, G.-Z. Pan, K. M. Fock, C. Fiocchi, S. K. Lam, and S.-D. Xiao, "Management consensus of inflammatory bowel disease for the asia?pacific region," *Journal of Gastroenterology and Hepatology*, vol. 21, no. 12, pp. 1772–1782, 2006.

[176] D. C. Baumgart, "The diagnosis and treatment of crohn's disease and ulcerative colitis," *Deutsches Aerzteblatt Int*, 2009.

[177] S. Kugathasan and C. Fiocchi, "Progress in basic inflammatory bowel disease research," *Seminars in Pediatric Surgery*, vol. 16, no. 3, pp. 146–153, 2007.

[178] J. D. Lewis, "The utility of biomarkers in the diagnosis and therapy of inflammatory bowel disease," *Gastroenterology*, vol. 140, no. 6, 1817–1826.e2, 2011.

[179] J. Tibble, "A simple method for assessing intestinal inflammation in crohn's disease," *Gut*, vol. 47, no. 4, pp. 506–513, 2000.

[180] D. Gevers, S. Kugathasan, L. A. Denson, Y. Vzquez-Baeza, W. Van Treuren, B. Ren, E. Schwager, D. Knights, S. J. Song, M. Yassour, X. C. Morgan, A. D. Kostic, C. Luo, A. Gonzlez, D. McDonald, Y. Haberman, T. Walters, S. Baker, J. Rosh, M. Stephens, M. Heyman, J. Markowitz, R. Baldassano, A. Griffiths, F. Sylvester, D. Mack, S. Kim, W. Crandall, J. Hyams, C. Huttenhower, R. Knight, and R. J. Xavier, "The treatment-naive microbiome in new-onset crohns disease," *Cell Host & Microbe*, vol. 15, no. 3, pp. 382–392, 2014.

[181] M. Rajili-Stojanovi, F. Shanahan, F. Guarner, and W. M. de Vos, "Phylogenetic analysis of dysbiosis in ulcerative colitis during remission," *Inflammatory Bowel Diseases*, vol. 19, no. 3, pp. 481–488, 2013.

[182] B. P. Willing, J. Dicksved, J. Halfvarson, A. F. Andersson, M. Lucio, Z. Zheng, G. Jrnerot, C. Tysk, J. K. Jansson, and L. Engstrand, "A pyrosequencing study in twins shows that gastrointestinal microbial profiles vary with inflammatory bowel disease phenotypes," *Gastroenterology*, vol. 139, no. 6, 1844–1854.e1, 2010.

[183] B. Lai, F. Wang, X. Wang, L. Duan, and H. Zhu, "Intemap: Integrated metagenomic assembly pipeline for ngs short reads," *BMC Bioinformatics*, vol. 16, no. 1, 2015.

[184] F. Guo, F. Ju, L. Cai, and T. Zhang, "Taxonomic precision of different hypervariable regions of 16s rrna gene and annotation methods for functional bacterial groups in biological wastewater treatment," *PloS One*, vol. 8, no. 10, e76185, 2013.

[185] A. Brady and S. L. Salzberg, "Phymm and phymmbl: Metagenomic phylogenetic classification with interpolated markov models," *Nature Methods*, vol. 6, no. 9, pp. 673–676, 2009.

[186] J. B. Ewaschuk, "Probiotics and prebiotics in chronic inflammatory bowel diseases," *World Journal of Gastroenterology*, vol. 12, no. 37, p. 5941, 2006.

[187] V. Sizaire, F. Nackers, E. Comte, and F. Portaels, "Mycobacterium ulcerans infection: Control, diagnosis, and treatment," *The Lancet Infectious Diseases*, vol. 6, no. 5, pp. 288–296, 2006.

[188] M. Stoyanova, I. Pavlina, P. Moncheva, and N. Bogatzevska, "Biodiversity and incidence ofburkholderiaspecies," *Biotechnology & Biotechnological Equipment*, vol. 21, no. 3, pp. 306–310, 2007.

[189] P. Brouqui, B. Davoust, S. Haddad, E. Vidor, and D. Raoult, "Serological evaluation of ehrlichia canis infections in military dogs in africa and reunion island," *Veterinary Microbiology*, vol. 26, no. 1-2, pp. 103–105, 1991.

[190] T. Bennur, A. R. Kumar, S. Zinjarde, and V. Javdekar, "Nocardiopsis species: Incidence, ecological roles and adaptations," *Microbiological Research*, vol. 174, pp. 33–47, 2015.

[191] D. Nagy-Szakal, E. B. Hollister, R. A. Luna, R. Szigeti, N. Tatevian, C. W. Smith, J. Versalovic, and R. Kellermayer, "Cellulose supplementation early in life ameliorates colitis in adult mice," *PloS One*, vol. 8, no. 2, e56685, 2013.

[192] E. Papa, M. Docktor, C. Smillie, S. Weber, S. P. Preheim, D. Gevers, G. Giannoukos, D. Ciulla, D. Tabbaa, J. Ingram, *et al.*, "Non-invasive mapping of the gastrointestinal microbiota identifies children with inflammatory bowel disease," *PloS One*, vol. 7, no. 6, 2012.

[193] D. Hyatt, P. F. LoCascio, L. J. Hauser, and E. C. Uberbacher, "Gene and translation initiation site prediction in metagenomic sequences," *Bioinformatics*, vol. 28, no. 17, pp. 2223–2230, 2012.

[194] Y. Liu, J. Guo, G. Hu, and H. Zhu, "Gene prediction in metagenomic fragments based on the svm algorithm," *BMC Bioinformatics*, vol. 14, no. Suppl 5, S12, 2013.

[195] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with bowtie 2," *Nature Methods*, vol. 9, no. 4, pp. 357–359, 2012.

[196] F. Wilcoxon, S. Katti, and R. A. Wilcox, "Critical values and probability levels for the wilcoxon rank sum test and the wilcoxon signed rank test," *Selected tables in mathematical statistics*, vol. 1, pp. 171–259, 1970.

[197] S. Altschul, "Gapped blast and psi-blast: A new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[198] R. L. Tatusov, "The cog database: A tool for genome-scale analysis of protein functions and evolution," *Nucleic Acids Research*, vol. 28, no. 1, pp. 33–36, 2000.

[199] P. M. Sagar, R. R. Dozois, and B. G. Wolff, "Long-term results of ileal pouch-anal anastomosis in patients with crohn's disease," *Diseases of the colon & rectum*, vol. 39, no. 8, pp. 893–898, 1996.

[200] H. Tilg, P. D. Cani, and E. A. Mayer, "Gut microbiome and liver diseases," *Gut*, vol. 65, no. 12, pp. 2035–2044, 2016.

[201] T. Bennike, S. Birkelund, A. Stensballe, and V. Andersen, "Biomarkers in inflammatory bowel diseases: Current status and proteomics identification strategies," *World J Gastroenterol*, vol. 20, no. 12, pp. 3231–44, 2014.

[202] C. Kasai, K. Sugimoto, I. Moritani, J. Tanaka, Y. Oya, H. Inoue, M. Tameda, K. Shiraki, M. Ito, Y. Takei, *et al.*, "Comparison of the gut microbiota composition between obese and non-obese individuals in a japanese population, as analyzed by terminal restriction fragment length polymorphism and next-generation sequencing," *BMC gastroenterology*, vol. 15, no. 1, p. 100, 2015.

[203] T. Shimizu, K. Ohtani, H. Hirakawa, K. Ohshima, A. Yamashita, T. Shiba, N. Ogasawara, M. Hattori, S. Kuhara, and H. Hayashi, "Complete genome sequence of clostridium perfringens, an anaerobic flesh-eater," *Proceedings of the National Academy of Sciences*, vol. 99, no. 2, pp. 996–1001, 2002.

[204] I. Mukhopadhya, R. Hansen, E. M. El-Omar, and G. L. Hold, "Ibdwhat role do proteobacteria play?" *Nature Reviews Gastroenterology & Hepatology*, vol. 9, no. 4, p. 219, 2012.

[205] P. J. Christie and J. P. Vogel, "Bacterial type iv secretion: Conjugation systems adapted to deliver effector molecules to host cells," *Trends in microbiology*, vol. 8, no. 8, pp. 354–360, 2000.

[206] A. Visconti, C. I. Le Roy, F. Rosa, N. Rossi, T. C. Martin, R. P. Mohney, W. Li, E. de Rinaldis, J. T. Bell, J. C. Venter, *et al.*, "Interplay between the human gut microbiome and host metabolism," *Nature Communications*, vol. 10, no. 1, pp. 1–10, 2019.