CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

Metagenomic Analysis of Human Gut Microbiome as a Tool Towards Non-invasive Biomarkers for Type II Diabetes Mellitus

by

Wajeeha Mujahid

A thesis submitted in partial fulfillment for the degree of Master of Science

in the

Faculty of Health and Life Sciences
Department of Bioinformatics and Biosciences

2024

I dedicate this thesis to my husband, parents, daughter and my teachers.

CERTIFICATE OF APPROVAL

Metagenomic Analysis of Human Gut Microbiome as a Tool Towards Non-invasive Biomarkers for Type II Diabetes Mellitus

by

Wajeeha Mujahid

(MBS223016)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr.Syed Babar Jamal	NUMS, Islamabad
(b)	Internal Examiner	Dr.Erum Dilshad	CUST, Islamabad
(c)	Supervisor	Dr.Syeda Marriam Bakhtiar	CUST, Islamabad

Dr.Syeda Marriam Bakhtiar
Thesis Supervisor
September, 2024

Dr. Syeda Marriam Bakhtiar
Head
Dept of Bioinformatics and Biosciences
September, 2024

Dr. Sahar Fazal
Dean
Faculty of Health and Life Sciences
September, 2024

Author's Declaration

I, **Wajeeha Mujahid** hereby state that my MS thesis titled "**Metagenomic Analysis of Human Gut Microbiome as a Tool Towards Non-invasive Biomarkers for Type II Diabetes Mellitus**" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.

(Wajeeha Mujahid)
Registration No: MBS223016

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled "**Metagenomic Analysis of Human Gut Microbiome as a Tool Towards Non-invasive Biomarkers for Type II Diabetes Mellitus**" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.

(Wajeeha Mujahid)
Registration No: MBS223016

Acknowledgement

In the name of Allah, the Most Gracious and the Most Merciful Alhamdulillah, all praises to Allah for giving me strength and for His blessings in completing my MS thesis. First, I would like to express my gratitude to Capital University of Science and Technology (CUST) Islamabad my Alma matter for providing me an opportunity to do MS Biosciences and achieving my goal to pursue higher studies. I would like to extend my heartfelt gratitude to my supervisor, **Dr. Syeda Marriam Bakhtiar**, Assistant Professor, CUST for her continuous support, guidance, and motivation throughout this journey. Her mentorship played a significant role in the timely completion of my study. I would like to acknowledge the contributions of all my teachers Dr.Shaukat Iqbal, Dr.Erum Dilshad, Dr. Arshia Amin, Dr.Sania Riaz, Dr.Sohail Ahmad Jan and Dr.M.Asad Anwar.

Furthermore, I am grateful to Dr. Sahar Fazal, Dean of the Department of Bioinformatics and Biosciences. Finally, I would also like to acknowledge my parents, my husband and my lovely daughter for all the support, patience, and motivation. This achievement would not have been possible without their encouragement and assistance.

(Wajeeha Mujahid)

Abstract

Type II Diabetes Mellitus is a chronic metabolic disorder that is caused by two primary factors: defective insulin secretion by pancreatic β cells and insulin resistance by the tissues. The prevalence and the incidence rate of Type II Diabetes Mellitus is continuously increasing throughout the world. This increasing incidence rate of the disease is probably caused by a few reasons, including the dysbiosis of the gut microbiota that is now known as a separate organ to control numerous physiological processes. Considering the significant role of gut microbiome dysbiosis in the onset of various diseases including Type II Diabetes Mellitus, multiple studies have been conducted to unveil this relationship by associating certain microbial signatures such as composition and abundance of the gut microbial communities with the pathophysiology of the disease. On the other hand, many meta-analyses have also been conducted to determine the microbial signatures that can be utilized as non-invasive diagnostic toolkits. Unfortunately, these studies either used an assembly based or a read based approach for taxonomic profiling of gut metagenomes. Additionally, these studies employed samples from specific ethnicities which limit the applicability of these findings on other populations due to genetic or technical differences. Furthermore, these studies reported signatures of general dysbiosis rather than disease specific biomarkers. This is one crucial reason for discrepancies in the microbial signatures of Type II Diabetes Mellitus despite the conduct of large number of gut biomarkers studies. Therefore, this study aimed at identification of biomarkers by gut metagenome profiling of Type II Diabetes Mellitus patients with a novel approach. Our approach utilized samples from multiple populations i.e. America and China and used an integrated approach i.e. assembly based as well as read based methods to perform the differential analysis between healthy control and Type II Diabetes Mellitus gut metagenomes. This study design ensures reproducibility and robustness as it has determined biomarkers by avoiding confounder effects, as well as the technical and genetic variations. Implication of such integrated biomarkers identification approaches will broaden the scope and applicability of the findings across the populations.

Contents

List of Figures

List of Tables

Abbreviations

ADA	American Diabetes Association
ASCVD	Atherosclerotic Cardiovascular Disease
BA	Bile Acids
BCAA	Branched Chain Amino Acids
BGCs	Biosynthetic Gene Clusters
BMI	Body Mass Index
CD	Crohn's Disease
DBG	De Bruijn Graph
DM	Diabetes Mellitus
FPG	Fasting Plasma Glucose
GALT	Gut Associated Lymphoid Tissue
GIT	Gastro-Intestinal Tract
GM	Gut Microbiota
HQ	High Quality
IBD	Inflammatory Bowel Disease
IDF	International Diabetes Federation
LQ	Low Quality
MAGs	Metagenome Assembled Genomes
MetS	Metabolic Syndrome
MQ	Medium Quality
NAFLD	Non-Alcoholic Fatty Liver Disease
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
OLC	Overlapping Layout Consensus

OGTT	Oral Glucose Tolerance Test
PG	Plasma Glucose
QC	Quality Control
rRNA	Ribosomal RNA
SCFA	Short Chain Fatty Acids
SNPs	Single Nucleotide Polymorphisms
SRA	Sequence Read Archive
T2DM	Type 2 Diabetes Mellitus
WHO	World Health Organization

Chapter 1

Introduction

1.1 Background

One of the oldest recognized illnesses, diabetes mellitus (DM), was first documented in Egyptian literature from about 3000 years ago [1]. It is a long-lasting health condition where your blood sugar levels are too high. This happens either because your body doesn't use insulin well or because it doesn't make enough insulin. There are two main types for this condition: Type I, which is autoimmune mediated, and Type II, which is caused by external stimuli [2]. Ninety percent of instances of diabetes are type 2 diabetes (T2DM), which is caused by a combination of decreased insulin production and insulin resistance [3].

According to 2019 reports, diabetes is the fifth biggest cause of mortality worldwide with a high prevalence [2]. The rapidly rising incidence of type 2 diabetes is probably caused by several reasons, including the aging of the population, the rising incidence of obesity, and the fall in physical activity levels that comes with industrialization. Risk factors for T2DM include being older, obese, having prior gestational diabetes, having a family history of the condition, having low glucose tolerance, being inactive, and being a certain race or ethnicity [4]. It is now acknowledged that the gut microbiota is a separate organ that controls numerous physiological processes and influences a wide range of host activities. The trillions of bacteria, the term "human microbiota" refers to the group of bacteria, fungi, and

eukaryotes that live in the respiratory tract, gastrointestinal tract, genitourinary tract, and skin of humans. microbiota. mucosa, and the mammary glands. One of the biggest interfaces (250–400 m^2) in the human body exists between the host, external stimuli, and antigens [5]. This interface is found in the gastrointestinal tract. Approximately 60 tons of food and a multitude of environmental microbes represent a serious threat to the integrity of the human gastrointestinal system throughout an average lifetime. The GI tract is thought to include around 10^{14} different microorganisms, including over 100 times the amount of genetic content (microbiome) found in human genomes and approximately ten times as many bacterial as human cells are present [6].

It is noteworthy here that there is a difference between the terms "microbiome" and "microbiota", as the former refers to the total genetic content of micro-organisms inhabiting a particular niche, whereas the latter signifies the total microbial community. These micro-organisms interact with the host in mutualistic, pathogenic, or commensalism mode. The microbiota offers the host many benefits through a range of physiological functions, such as maintaining gut integrity or obtaining energy, preventing infections, and regulating the host's immunity [7]. These host-microbe interactions, which have important functions in the body, maintain human health including immune regulation, synthesis of vitamins (B12 and K), digestion of food, and metabolization of xenobiotic. These microbial communities can be perturbed by external factors such as use of antibiotics, age, diet, pregnancy, stress etc. and undergo a change in either their composition or abundance state known as dysbiosis [8]. The state of dysbiosis can increase various disease susceptibilities and presentation of different phenotypes including neurological, immunological, physiological, psychic, and metabolic disorders. One such metabolic disorder caused by the dysbiosis of the microbiome of the human gut is T2DM [9]. Previous studies show biomarkers which are not specific to T2DM. With the advent of NGS technology, there has been a rapid surge in metagenomic studies substantially advancing our knowledge of unculturable microbial majority and their association with prevalent diseases and ecological changes [10]. The conventional approach that is adopted to discover disease specific biomarkers from metagenomics data often involves a differential method [9]. However, in

many cases, biomarkers identified using this differential approach reflect a mix of disease-specific characteristics and the features of general dysbiosis [11].

1.2 Global prevalence of T2DM

Diabetes is the sixth most common cause of death worldwide; with a significant prevalence The International Diabetes Federation estimates that 463 million people between the ages of 20 and 79 have diabetes in 2019 and that number is projected to rise to 700 million by 2045. The IDF states that type 2 diabetes affects more than 90% of diabetic individuals. 11.6% of Chinese adults are thought to have diabetes. 13% of Americans who are eighteen years of age or older estimated to have diabetes. Premature mortality is highly risked by diabetes. One person is thought to pass away every eight seconds. Because of androgen excess in women and androgen deficiency in men, it is popular in both sexes. This high prevalence is still an underestimation due to the inefficient reporting systems in many parts of the world including Pakistan and is concerning enough to require immediate attention [12].

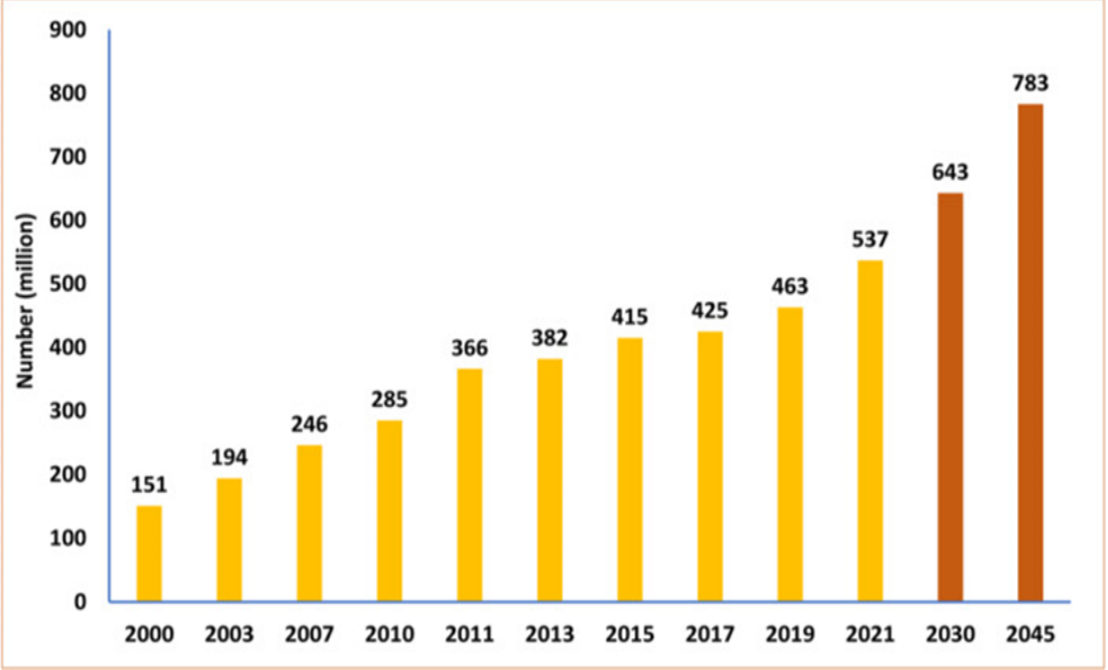


FIGURE 1.1: Estimated number of cases of diabetes in 2030 to 2045

1.3 Pathophysiology of T2DM

T2DM is the cause of over 90% of diabetes occurrences globally deficiencies in the pancreatic beta cells' ability to synthesize and release insulin as well as in the way that muscle, adipose tissue, and/or liver cells react to insulin might cause it to occur. The generation, release, and uptake of insulin by target tissues are all governed by intricate systems, and malfunctions in any of these processes can result in metabolic dysregulation and the eventual beginning of T2DM. For instance, deficiencies in any of the insulin binding pathways may result in decreased glucose absorption, raising blood glucose levels (hyperglycemia) and putting more strain on the pancreatic beta cells that produce insulin [13].

1.3.1 Role of Microbiota in Human Growth

The assumption that the prenatal gastrointestinal environment is sterile stems from the belief that the placenta protects the fetus from any bacteria that could jeopardize its survival. This theory's supporting evidence comes from the idea that any bacterium in the uterus is dangerous for the growing fetus and is associated with deformities and preterm birth. The way that a baby is fed, the type of birth, and pregnancy all influence the newborn microbiome. Early-life microbiota acquisition affects the maturation of the immune system and its potential to either promote or inhibit disease in the future [14]. *Lactobacillus spp.*, *Bifidobacterium*, and *Prevotella* are among the vaginal bacteria that colonize neonates delivered vaginally, whereas the mother's skin microbiota, *Corynebacterium*, *Staphylococcus*, and *Propionibacterium spp.*, colonize newborns delivered by caesarean surgery. When solid food is introduced and weaning occurs, the diversity of the gut microbiota increases, which benefits butyrate-producing bacteria like Clostridium [15].

1.3.2 Role of Microbiota in Immunity

The immune system is made up of an intricate network of innate and adaptive cells that may respond to a variety of stimuli, preserve tissue, and repair it when

it encounters microorganisms. Mucus, immune cells, antimicrobial peptides, and IgA work in remarkable concert in the human gut to form what is known as the "mucosal firewall." To maintain intestinal homeostasis, this synergy is crucial in stopping bacteria from producing inflammation by allowing them to pass through the lamina. Antibiotic use, age, diet, and lifestyle all have an impact on the gut flora. Numerous illnesses are caused by this disturbance of intestinal homeostasis [7].

1.3.3 Metabolic Diseases and Microbiome

It is acknowledged that the gut microbiota is a metabolically active "organ," due to its distinct and dynamic environment. Thus, immune system and microbial interactions control gastrointestinal homeostasis. However, many disorders, including obesity and T2DM, are brought on by a breakdown of gastrointestinal homeostasis [7].

1.3.4 Gut Microbiota Influence on Glucose and Insulin Metabolism

A few ways that the gut microbiota can impact host glucose homeostasis are through the fermentation process, which produces metabolites and their effects, the activation of inflammatory cascades that release cytokines, the disruption of the permeability of the intestinal mucosal barrier, which permits toxins to enter the body, and direct signaling through incretin secretion.

The membrane transport of sugars, branched chain amino acids (BCAAs), methane metabolism, xenobiotic metabolism, and sulphate reduction improves in type 2 diabetic patients. The same group showed reductions in flagellar assembly, butyrate synthesis, cofactor and vitamin metabolism, and bacterial chemotaxis [16].

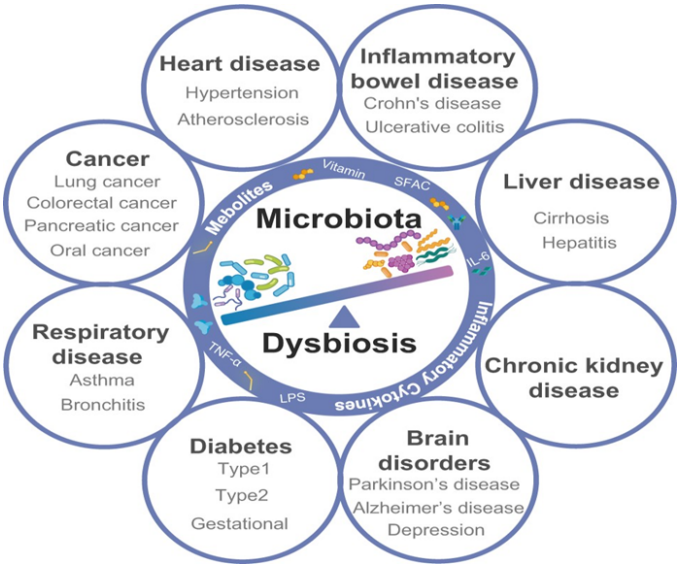


FIGURE 1.2: Gut microbiota dysbiosis leads to several diseases

1.4 Available Diagnostic Biomarkers for T2DM

To identify T2DM biomarkers, a lot of research has been conducted. Many evaluations have identified genetic and non-genetic susceptibility/risk biomarkers for the development of T2DM. Often referred to as "genomics," genome science is the study of an organism's complete DNA sequence, including its composition, expression, and biological roles. Genetic variants that are examined by genomics include single nucleotide polymorphisms (SNPs) and chromosomal abnormalities linked to illnesses. T2DM genomics study used DNA microarray, next-generation sequencing, and TaqMan qPCR assays [17]. Another study that compiled omics-based biomarkers for tracking different types of diabetes used non-invasive biomarkers from saliva and urine. Fecal samples have also been used in genetic investigations related to T2DM. Although minor amounts of human DNA can be found, most of the genetic material found in fecal samples is bacterial. Human DNA extraction from feces samples is being improved, but there are still many obstacles to overcome before it can be used in population research. Thus, all T2DM research using fecal samples is metagenomics research [18].

1.5 Human Gut Microbiome Studies and Associations With T2DM

The swift advancement of next-generation sequencing (NGS) technologies and bioinformatics techniques for data processing has ushered in the field of metagenome research, which examine the microbial makeup of natural populations. With the advent of NGS technology, there has been a rapid surge in metagenomic studies substantially advancing our knowledge of unculturable microbial majority and their association with prevalent diseases and ecological changes [9]. Facilitating the economical high-throughput sequencing of millions of DNA fragments NGS has emerged as a preferred tool for diagnostics and precision medicine applications, allowing simultaneous analysis of multiple biomarkers within a single sample. Improvements in the investigation of the intestinal microbiome over the past ten years have shown Since alterations in the composition of gut bacteria are connected to a number of diseases, including diabetes and obesity. There is ongoing debate on the best way to use shotgun sequencing or 16S RNA profiling, how to best cover the V1–V9 hypervariable regions differently, how to more precisely quantify the components of the microbiota, and how to establish universal sample acquisition methodologies [19].

1.5.1 Metagenomics As A Powerful Method for Determining Microbial Diversity

The word metagenome denotes a group of genomes found in samples that are being studied for functional analysis and cloning. The advancement of whole genome and 16S rRNA-based microbial investigations has been expedited by the introduction of sequencing technology. Studies looking at the genomic information of microorganisms have been referred to as "metagenomics," which is further classified into two categories: amplicon and shotgun metagenomics. Studies on amplicon metagenomics generally investigate the variety of microorganisms, whereas research on shotgun metagenomics mostly concentrates on identifying functional

genes and metabolisms. Like amplicon metagenomics, shotgun metagenomics results also include taxonomic diversity and show several microbial metabolisms that may be deduced from whole genome sequencing. Genomic investigations linked with shotgun metagenomics have demonstrated a strong potential for discovering novel bacteria and viruses and forecasting the primary metabolisms in the habitats under investigation. The four basic phases of metagenomics include sampling and DNA extraction, sequencing, analysis, and visualization, even though the data are different. (Figure 1.3). Metagenomic studies frequently consist of a continuous series of processes where the results of one step determine the next. Since samples for metagenomics study are taken straight from the field, it is important to avoid cross-contamination [10, 20].

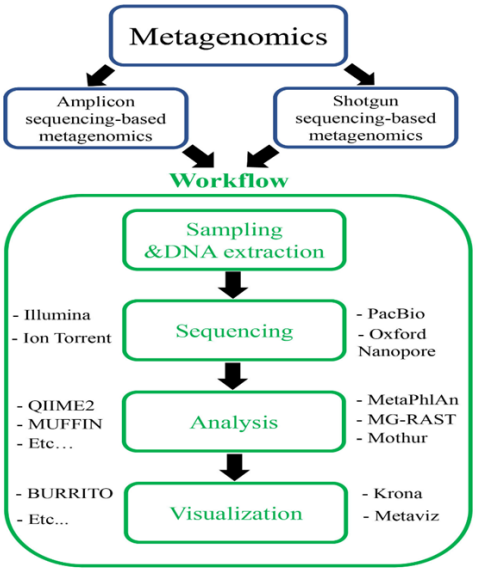


FIGURE 1.3: Metagenomics workflow using different sequencing platforms and bioinformatics tools.

1.5.2 Illustration of False Positive and False Negative Results

One of the most effective and widely applicable methods for integrating genomic and molecular data into clinical practice is biomarker discovery. Human microbial communities can function as biomarkers for host characteristics like lifestyle and illness, according to recent metagenomic tests.

These results have also brought attention to the importance of tasks such as class prediction, which determines the subtype of a new sample, and class discovery, which finds new subtypes of a disease. These findings are supported by comparisons between healthy and diseased tissues [20].

The conventional approach that is adopted to discover disease specific biomarkers from metagenomics data often involves a differential method. This method involves comparison of case and control metagenomic samples to identify differences in the taxonomic composition of microorganisms, their abundance or the gene and protein expression [8]. However, in many cases, biomarkers identified using this differential approach reflect a mix of disease-specific characteristics and the features of general dysbiosis. As a result, when such biomarkers are evaluated across different datasets, they illustrate a weak predictive power [10].

A critical evaluation of such studies reflects an immediate need to improve the current approaches adopted to evaluate the predicted diagnostics biomarkers so that the general dysbiosis features can be separated from disease-specific biomarkers. By acknowledging the existing limitations of conventional study designs that often lead to the false discovery of disease-specific biomarkers and re-discovering the diagnostic biomarkers for critical diseases such as diabetes, this study has the potential to pave the way for more accurate diagnostic and targeted interventions. Therefore, in this study we will design a cross-validation approach by extending the concept of "subtractive metagenomics" with the potential to accurately identify disease-specific biomarkers.

1.6 Problem Statement

Current metagenomics approaches for identifying T2DM biomarkers primarily use either read-based or assembly-based methods. Read-based methods are typically employed to assess taxonomic abundance or microbial composition, while assembly-based methods are used to identify functional biomarkers. However, these studies are often conducted on samples from a single population, resulting

in findings that are specific to that population's genetics and not broadly applicable. Therefore, there is a need for a revised methodology, such as an integrative approach, which combines both methods. Additionally, future studies should utilize samples from multiple populations to focus on disease-specific rather than population-specific biomarkers.

1.7 Aim and Objectives

This study aims to identify non-invasive biomarkers of Type II Diabetes Mellitus. Our specific objectives will include:

1. To determine the number and taxonomic makeup of the human gut microbiome in individuals with Type II Diabetes Mellitus by comparing them with healthy control groups through datasets which are publicly available.

2. To explore the taxonomic diversity within the Type II diabetes Mellitus specific metagenomic samples.

3. To perform differential analysis of taxonomic profiles of healthy and Type II Diabetes Mellitus gut metagenomes to predict Type II Diabetes Mellitus biomarkers.

Chapter 2

Literature Review

An overview of the body of research on the worrying prevalence of T2DM is given in this section. its classification by WHO, disease pathogenicity, available treatments conventional approach that is commonly used for biomarkers discover, challenges faced by the discovery process, and available computational tools. Additionally, it also provides a glimpse of existing research concerning T2DM and microbiome associations.

2.1 T2DM

When therapy is not obtained, a group of metabolic diseases collectively referred to as "diabetes" are characterized by the presence of hyperglycemia. The varied etiopathology include abnormalities in protein, lipid, and carbohydrate metabolism in addition to deficiencies in insulin action, secretion, or both. Diabetes can have long-term effects, such as neuropathy, nephropathy, and retinopathy. Common signs of diabetes include impaired vision, thirst, polyuria, and weight loss. A non-kenotic hyperosmolar condition or ketoacidosis are the most dangerous clinical indicators since they can lead to treatment failure, dehydration, and even death [21].

T2DM accounts for 90% of cases of diabetes and is brought on by a confluence of insulin resistance and decreased insulin production [3]. T2DM is becoming more widespread around the world. In the next ten years, the disease is predicted to affect twice as many people due to an aging population and as a result it is fast spreading to other parts of the world. This will increase the already demanding burden for medical practitioners, particularly in underdeveloped nations. Screening and diagnosis are still based on guidelines from the American Diabetes Association (ADA) and the World Health Organization (WHO) which consider both laboratory and clinical findings. Treatment options for the illness include lifestyle modifications, weight management, oral hypoglycemic medicines, and the use of insulin sensitizers such as metformin, a biguanide that lowers insulin resistance. However, there is presently no known cure for the disorder. The first-line drug that is still recommended is metformin, especially for patients who are obese [1].

2.1.1 Global Burden and Epidemiology of Diabetes

Diabetes affects people everywhere, but it particularly affects those who live in rural areas of low- and middle-income countries. In 2014, 422 million individuals globally were estimated by the World Health Organization to have diabetes. The proportion of people living with diabetes is steadily rising. Furthermore, 1.1 million children and adolescents between the ages of 14 and 19 are estimated to have type 1 diabetes by the International Diabetes Federation (IDF). By 2045, at least 629 million people worldwide will suffer from diabetes if nothing is done to stop the epidemic's rising prevalence [22]. The most recent estimates indicate that the prevalence of diabetes in the Caribbean and Northern America was 11.1% in 2019 and is expected to rise to 13% by 2045. The largest incidence rates are found in the Middle East and North Africa, where an additional 13.9% of cases are predicted by 2045. Africa now has the lowest prevalence rate (4.7%), Increases reaching 5.2% are projected by 2045. High or intermediate frequencies are found in most countries in South America and Southeast Asia. 463 million people worldwide have diabetes, according to a 2019 study by Saeedi et al. which translates to a prevalence rate of 9.3%. The anticipated rise in this prevalence

rate is 10.2% and 10.9%, by 2030 and 2045, respectively. The diabetes prevalence, stratified by region calculations for some of those countries show which countries have the highest number of diabetic patients worldwide in 2019. Several countries' region-stratified diabetes prevalence is calculated, presenting the countries with the largest worldwide diabetes patient populations in 2019 [4].

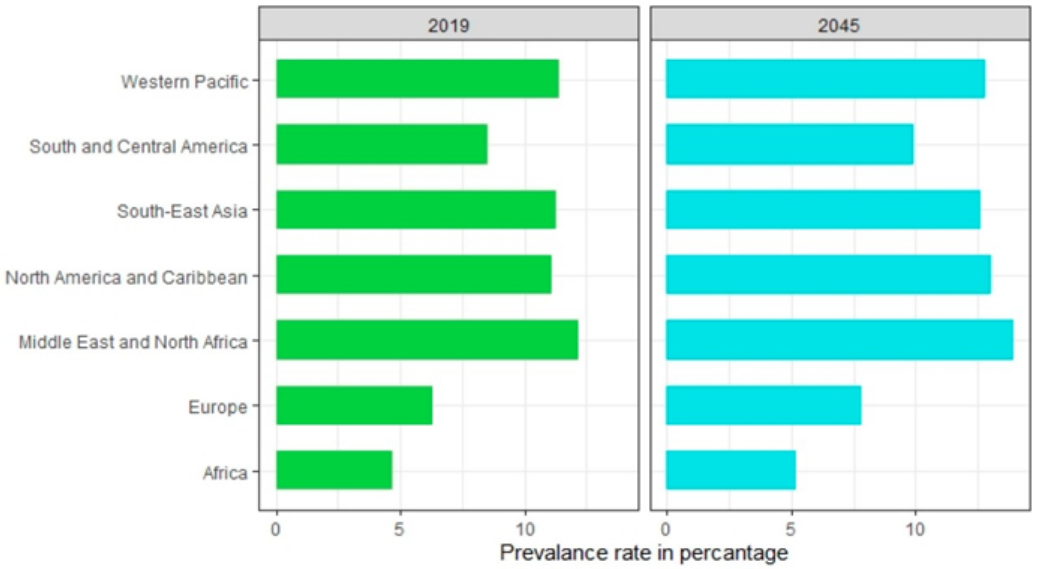


FIGURE 2.1: Global diabetes prevalence in 2019 and 2045 (estimated).

The top ten nations or territories with the highest rate of diabetes prevalence are listed in figure. China, with about 116 million diabetes patients, has the highest percentage of these individuals among the other countries. In the future decade, the United States of America is expected to be among the nations where diabetes risk is greatest, with 31 million people, trailing behind India, which comes in second on the list with 77 million. According to predictions, Pakistan, Brazil, and Mexico will have high-end diabetic populations, with approximately There are, correspondingly, Worldwide, there are 19, 16, and 12 million diabetics. Bangladesh is at the bottom of this scale, although it has an equivalent risk of diabetes as the US due to its growing population and lack of well-thought-out intervention measures [4].

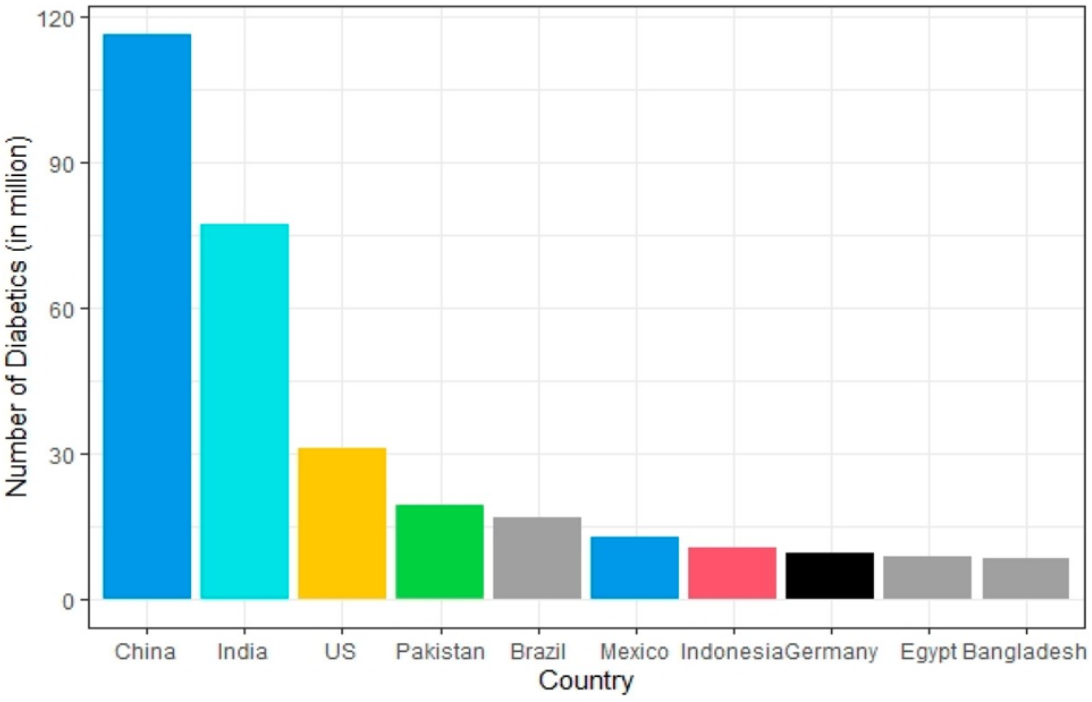


FIGURE 2.2: Nations with the greatest global concentration of diabetic patients in 2019.

It has far-reaching socioeconomic ramifications and poses a threat to economies and productivity at the national level, especially in nations with lower and moderate incomes where diabetes and other illnesses frequently coexist [4].

2.1.2 Diabetes Aetio-pathology

Nowadays, it is commonly acknowledged that the primary cause of pancreatic β-cell dysfunction or loss is a feature shared by all types of diabetes. Numerous pathways may result in β-cells losing their ability to function or dying entirely. These pathways include insulin resistance, autoimmunity, concomitant diseases, inflammation, genetic predisposition and anomalies, and epigenetic processes. Distinguishing between Problems with β-cells and decreasing the mass of βcells may have important consequences for therapeutic strategies aimed at preserving or enhancing glucose tolerance. Knowing the β-cell status can assist identify diabetes subtypes and direct treatment [23].

2.1.3 Pathogenesis of T2DM

There are multiple distinct pathophysiologic problems associated with T2DM. Increased endogenous glucose production and decreased peripheral glucose absorption are recognized characteristics of insulin resistance. especially in muscle. The accumulation of intermediate lipid metabolites increased peripheral glucose use, and accelerated lipolysis culminated in increased glucose output, decreased beta-cell function, and impaired peripheral glucose consumption. It has been established that the development of type 2 diabetes is significantly influenced by inflammation and insulinocyte insulin resistance. It is believed that there is a strong correlation between the prevalence of non-alcoholic fatty liver disease (NAFLD) and insulin resistance. The terms ”glucotoxicity” and ”lipotoxicity,” initially described how beta cells degrade due to extended increases in fat and glucose levels (also known as ”nutritoxicity”), which now affect all nutrients. Beta cell activity is already abnormal and will further worsen over time, even if the compensatory insulin secreted by the pancreatic beta cells may initially keep plasma glucose levels within permissible ranges. The flora in the gut may have an impact on the hormonal and metabolic changes linked to T2DM [24].

2.1.4 Type 2 Diabetes Mellitus Site of Insulin Resistance

Both a normal insulin secretory response by the pancreatic beta cells and proper tissue sensitivity to the distinct effects of hyperglycemia and hyperinsulinemia (i.e. the mass-action effect of glucose) are necessary to maintain glucose homeostasis. Therefore, for the combined effects of insulin and hyperglycemia to promote glucose excretion, three intricately connected pathways must exist [24]:

1. Inhibition of the body’s normal manufacture of glucose, primarily in the liver.
2. Induction of the splanchnic tissues’ absorption of glucose.
3. Boosting peripheral tissues, especially muscles, to absorb glucose.

2.1.5 The Ominous Octet

Two features of insulin resistance are known to exist: a decrease in peripheral glucose absorption and an increase in endogenous (hepatic) glucose production. The accumulation of intermediate lipid metabolites leads to enhanced lipolysis, which promotes glucose generation even while peripheral consumption declines. After reaching a maximum, the compensatory insulin released by the beta cells in the pancreas progressively diminishes. Moreover, the pancreatic alpha-cells emit aberrant amounts of glucagon, mainly during the postprandial phase.

A contributing factor to both impaired insulin and excessive glucagon secretion in type 2 diabetes has been suggested to be the ”incretin defect,” which is essentially defined as an inadequate response of the gastrointestinal ”incretin” hormones to meal ingestion in addition to islet-cell resistance to the potentiating action on insulin-secretion by these gastrointestinal peptides. Furthermore, the body’s capacity to use circulating insulin to lower glucose production is hampered by the central nervous system, which is commonly observed in people with type 2 diabetes. This suggests that we should adopt the idea of the ”omnious octet” in favor of the ”triumvirate.” When treating hyperglycemia in T2DM patients, several pathogenic pathways must be considered. Figure 2.3 lists the eight major recognized risk factors for hyperglycemia brought on by the onset of type 2 diabetes [24].

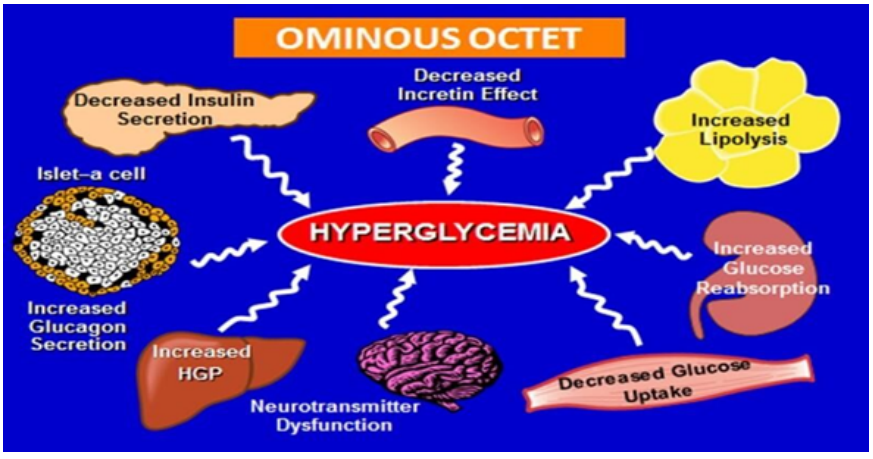


FIGURE 2.3: Eight Main Processes Underlying Hyperglycemia in T2DM Patients.

2.2 The Gut Microbiome

Recent research has revealed that several metabolic disorders may be influenced by the gut microbiome. Variations in the gut microbiota's composition in high-risk patients may act as early warning signs of T2DM.

Butyrate is one of the byproducts of intestinal bacteria that can have positive metabolic effects by promoting mitochondrial activity, preventing metabolic endotoxemia and stimulating intestinal gluconeogenesis through various pathways of hormone control and gene expression. Much effort is currently being put into understanding among other things whether the effects of the gut bacteria that produce butyrate and other bacterial products are the same, and their ultimate goal is to pave the path for more potent treatments for T2DM and obesity [25].

2.3 Diagnosis

Diabetes can have serious consequences if left untreated, thus early detection can help prevent these consequences. The primary signs of diabetes include persistently elevated blood glucose, increased thirst, increased hunger, and frequent urination.

It is common practice to conduct specific biochemical tests to detect diabetes or prediabetes. Diabetes is commonly diagnosed using oral glucose tolerance test (OGTT) and glycosylated hemoglobin [26].

2.3.1 Fasting Plasma Glucose

After an eight-hour fast, a blood sample is taken over night. According to the American Diabetes Association, a fasting plasma glucose (FPG) level of greater than 126 mg/dL (7.0 mm/L) is compatible with the diagnosis [26].

2.3.2 Oral Glucose Tolerance Test

The plasma glucose level is measured both before and two hours after consuming 75 grams of glucose. Diabetes is suspected if the 2-hour sample's plasma glucose (PG) level is more than 200 mg/dL (11.1 mmol/L). It is also a standard test, but in addition to having serious variability issues, it is more costly and inconvenient than FPG. Patients are required to consume a minimum of 150 grams of carbohydrates per day for three to five days. They are also urged not to take any medications that may complicate glucose regulation, such as steroids and thiazide diuretics [26].

2.3.3 Glycated Hemoglobin (Hb) A1c

Diabetes is also commonly diagnosed with the HbA1c test. The results of this test show the average blood glucose levels for the two to three months before. When a patient's HbA1c is greater than 6.5% (48 mmol/mol), diabetes mellitus (DM) is diagnosed. The Hb A1C test is quick, accurate, and uniform due to pre-analytical parameters. HbA1c has several problems besides being expensive, such as decreased sensitivity. Many conditions, including hemodialysis, pregnancy, sickle cell disease, blood transfusions or loss, and erythropoietin therapy, might cause it to alter [26].

2.4 Prognosis of T2DM

Given the elevated risk of both DM and atherosclerotic cardiovascular disease (ASCVD), lowering blood pressure, taking statins, exercising frequently, and giving up smoking are all important risk reduction strategies. Individuals have a 15% increased excess mortality risk on average when they have type 2 diabetes, though this varies greatly. In the United States, 4.4% of adult diabetics have vision-threatening diabetic retinopathy, while 1% of people have end-stage renal disease [26].

2.5 Role of Gut Microbiota in T2DM

The microbiota that coexists with our bodies greatly affects human health. Trillions of bacteria, fungi, archaea, and eukaryotes that live in the skin, gastrointestinal system, genitourinary tract, respiratory tract, mucosa, and mammary glands are together referred to as the human microbiota. The GI tract is thought to include around 1014 different microorganisms, including over 100 times the amount of genetic content (microbiome) found in human genomes and approximately 10 times the number of bacterial cells compared to human cells. A crucial component in preserving the gastrointestinal tract's (GIT) equilibrium is the intricate ecology known as the human gut microbiota. It is now acknowledged that the gut microbiota is a separate organ that controls numerous physiological processes and influences a wide range of host activities. Bacterial dysbiosis has been connected to diseases like obesity, both inflammatory bowel disease and type 2 diabetes. According to recent studies, demonstrated the role of gut microbiota in immune system diseases [6].

2.5.1 Bacteria Involved in T2DM

The strongest and busiest connection between our immune system and both advantageous and dangerous bacteria is found in the gastrointestinal tract. A wide variety of gut microbiota (GM) consists of commensal microorganisms like viruses, bacteria, fungi, and protozoa. These bacteria get their energy from the substrates that food supplies and coexist in a complicated balance with their host. There are the same number of bacteria 3 8.10_{13} to be exact that reside in human cells [27]. According to the latest 42 observational studies examining T2DM and the bacterial microbiome in humans, most of the research revealed associations between certain taxa and the disease or its symptoms. Together with *Proteobacteria* and *Actinobacteria*, gram-positive *Firmicutes* and gram-negative *Bacteroidetes* contribute for 60 – 80% and 20 – 30% of total GM, respectively. T2DM was inversely correlated with the *Ruminococcus*, *Fusobacterium*, *Blautia*, and *Bacteroides* genera, and positively linked with the genera of *Faecalibacterium*, *Akkermansia*, *Roseburia*,

and *Ruminococcus* according to the common and consistent data. It seems that microorganisms in the genus *Bifidobacterium* may offer protection against T2DM. Nearly every publication reports a negative correlation between this species and T2DM. Nevertheless, enhanced glucose tolerance was shown in almost all animal research on *Bifidum*, *B. ifantis*, *B.animalis*, *B.pseudocatenulatum*, *B.breve* that evaluated many species from this genus.

Thus, evidence from animal studies lends credence to the idea that *Bifidobacterium*, which is present in the human gut naturally or can be taken as probiotic supplements, protects against type 2 diabetes. *Bacteroides* was the genus that is second most recorded. The prevalence of this genus and type 2 diabetes have been linked in eight studies. Throughout five case-control studies, the T2DM group had fewer incidences of *Roseburia* than the healthy controls. According to two case-control investigations, the disease group's *Faecalibacterium* frequencies were lower within the commensal microbiome *Akkermansia muciniphila* is a relatively recent addition.

It was discovered in animal models to have a beneficial effect on the host's metabolism of glucose. Human studies have shown a negative link between the number of this bacterium and T2DM, which is consistent with findings on animals [28].

2.5.2 Potential Mechanisms of Microbiota Effects on Meta -bolism in T2DM Patients

There are numerous biological pathways linking T2DM and metabolic disorders to the gut microbiome. In the mammalian host, the microbiota impacts insulin sensitivity, glucose and lipid metabolism, intestinal permeability, and general energy homeostasis. It also interacts with food ingredients and modifies inflammation [28].

2.5.3 Gut Microbiota Impact on Glucose and Insulin Meta -bolism

Through a multitude of mechanisms, including the production of metabolites during fermentation and their subsequent secondary effects, the activation of inflammatory cascades leading to the release of cytokines, the disruption of the intestinal mucosal barrier's permeability, which permits toxins to enter the body, and direct signaling through incretin secretion gut microbiota can affect host glucose homeostasis. Additionally, the generation of gut hormones that control this process and the digestion of sugars, as well as essential metabolic organs including the liver, muscle, and clean fat, can all be impacted by gut bacteria and type 2 diabetes. One potential probiotic, *Bifidobacterium lactis*, for example could boost the synthesis of glycogen and reduce the expression of hepatic gluconeogenesis - related genes. Sugar transport across membranes, BCAA transport, methane metabolism, xenobiotic breakdown and metabolism, and reduction of sulphate are all enhanced in T2DM patients. The same group showed decreased levels of vitamin and cofactor metabolism, bacterium chemotaxis, filament assembly, and butyrate production [16].

2.5.4 Increased Gut Permeability

One trait of people with T2DM is increased intestinal permeability. Endotoxemia from metabolic processes and the translocation of gut microbial metabolites into the circulation are the outcomes. T2DM metabolic syndrome (MetS), obesity, and other metabolic illnesses have been demonstrated to be protected against by a variety of bacterial species present in the microbiome. The synthesis of water-soluble vitamins, protein catabolism, and secondary bile acids (BAs)3 are all aided by the GM. Its significance is also growing because of its function in gutassociated lymphoid tissue (GALT), innate and adaptive immunity formation and maintenance, and the process of deriving energy from otherwise indigestible meals [13].

2.5.5 Role of Microbiota in the Effectiveness of T2DM Drug Therapy

The local microbiota has an impact on the therapeutic efficacy and possible adverse consequences after antidiabetic medication administration. Antibiotics, non-antibiotic drugs, and anti-diabetic drugs are well known for their ability to modify microbiota and worsen diabetes. Like this, the baseline microbiota can influence many chemicals through a range of mechanisms and the pharmacokinetics and pharmacodynamics of pharmaceuticals. A comprehensive examination of 271 oral medications revealed that at least one bacterial strain is responsible for 66% of their metabolism. Thus, comprehending this reciprocal relationship and its impact on the clinical results of anti-diabetic medications could lead to the creation of novel approaches for the control of diabetes type 2. Research conducted on mice has demonstrated that GM regulates glucose homeostasis and satiety via the incretin hormone GLP-1, which is secreted by intestinal L cells. . Accordingly, a type of antidiabetic medication called GLP-1 receptor agonists can alter the Firmicutes to Bacterodites ratio, changing the composition of GM. DPP-4 inhibitors lower blood sugar by preventing GLP-1 from being broken down. They also restored the GM composition, which increased the number of Bacteroidetes. Combining sitagliptin with pre- and probiotics was successful in lowering several T2DM parameters. As a result, research on microbiome has taken a new turn, concentrating on the relationship between microbiota and anti-diabetic medications [28].

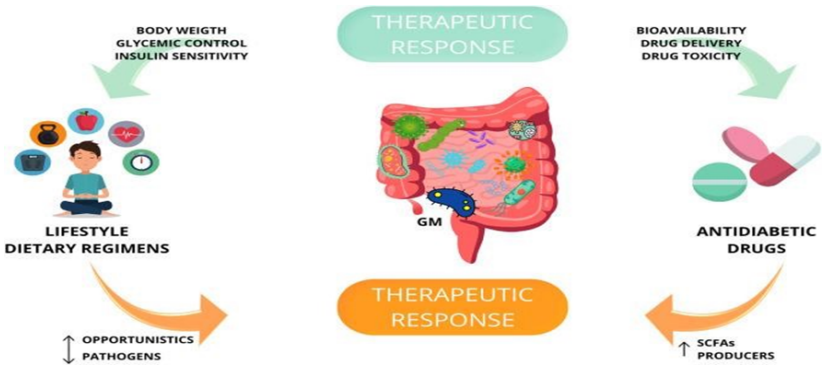


FIGURE 2.4: The role of microbiota in the therapeutic response.

2.6 Gut Microbiota as a Prognostic Biomarker or Diagnostic Predictor

Taxonomic analysis revealed a list of microbial risk markers, including moderate dysbiosis, a reduction in butyrate producers, an increase in a range of opportunistic pathogens, and an enrichment of microbial functions related to resistance to oxidative stress and sulphate reduction. Additionally, a specific gut metagenomic linkage group was identified and linked to the risk of T2DM. Additionally, an abundance of evidence suggests that inter-individual heterogeneity in the GM may predict the onset of disease and account for the range of responses to dietary plans.

Extensive research on precision nutrition indicates that dietary guidelines ought to be tailored to each patient's specific GM composition, and that genetic changes could account for the observed inter-individual heterogeneity in postprandial blood numerous cardiometabolic blood markers, including the inflammatory index, lipemic index, and postprandial glycaemic index as well as fasting and postprandial glycaemic indices, seem to be predicted by the composition of the total microbiome. It has been demonstrated that several microorganisms, including *Blastocystis spp.* and *Prevotella copri* are indicators of a healthy postprandial glucose metabolism. It's interesting to note that obese patients with higher abundances of *Akkermansia muciphila* appear to have higher insulin sensitivity when under calorie limitation; also, insulin sensitivity has been shown to improve in obese human subjects when treated with *Parabacteroides distasonis*.

Even when people adopt similar lifestyle choices, studying GM appears to help explain the variable risk of onset of metabolic illnesses. However, larger-scale research is increasingly needed to hypothesize personalized therapy for individual patients. Utilizing the microbiota as an adjuvant method can improve patients' reactivity to pharmacological treatment, improve adherence from patients, and help dysmetabolic individuals regain eubiosis by decreasing the side effects of their drugs [27].

2.7 T2DM & the Human Gut Microbiome: NGS Studies and Correlations

The rapid advancement of next-generation sequencing (NGS) technologies and bioinformatic data processing techniques has made it possible to conduct genome studies, which examine the microbial composition of natural inhabitants. Next-generation sequencing (NGS) technology's arrival has increased the quantity of metagenomic research has rapidly increased, significantly expanding our understanding of the majority of nonculturable bacteria and their relationships to common diseases and environmental shifts. With its ability to efficiently sequence millions of DNA fragments at high throughput, next-generation sequencing (NGS) has become the preferred approach for precision medicine and diagnostics, enabling the simultaneous study of several biomarkers in one sample. Over the past ten years, Developments in the field of intestinal microbiome research have shown a link between changes in the make-up of gut bacteria and several diseases, such as obesity and diabetes. The genetic causes of T2DM can be extensively studied using next-generation sequencing (NGS) technologies. These applications include:

1. Identifying common and rare genetic variants linked to the disease.

2. Conducting functional studies to elucidate the role of genes in disease pathogenesis.

3. Assessing the environmental factors contributing to the disease through the application of microbiome profiling techniques.

There is continuous discussion on the most effective ways to cover the V1–V9 hypervariable regions, quantify the microbiota components more precisely, implement universal sample acquisition procedures, and apply shotgun sequencing or 16S RNA profiling [10].

2.8 Metagenomics: An Effective Technique to Assess Microbial Diversity

The "omics" period has given rise to several sciences, including metagenomics, transcriptomics, proteomics, metabolomics, phenomics, and genomics. Metagenomics has been one of the key enablers of a notable surge in discoveries concerning the microbial realm. The diversity and roles of microorganisms on Earth are meaningfully revealed by recently found microbiomes in various ecologies. Consequently, novel microbe-based applications in the food business, agriculture, and human health have been made possible by the findings of metagenomic research.

Studies looking at the genomic information of microorganisms have been referred to as "metagenomics," which is further classified into two categories: amplicon and shotgun metagenomics. Shotgun metagenomics research mostly focuses on identifying functional genes and metabolisms, whereas amplified metagenomics studies generally investigate microbial diversity. Genomic investigations linked with shotgun metagenomics have demonstrated a strong potential for discovering novel bacteria and viruses and forecasting the primary metabolisms in the habitats under investigation. The four basic phases of metagenomics include sampling and DNA extraction, sequencing, analysis, and visualization, even though the data are diverse [29].

2.8.1 Shotgun Metagenomic Data Analysis

For metagenomics, there are various techniques for performing shotgun data analysis. Finding and eliminating contaminants and low-quality sequences using a range of computational approaches is a standard first step in quality control. Among them are programs like Cut adapt and FastQC. Fast Screen compares the reads to several reference genomes, including yeast, Escherichia coli, human, and mouse, to produce a succinct summary of the alignments between the reads. The reads can be sent straight to taxonomic classifiers or joined into longer contiguous sequences or contigs after quality control. Because taxonomic categorization places read

into bins according to their taxon ID, it is a type of binning. Other features, like composition and co-abundance profiles can also be used for binning however these methods usually need the assembly of reads into longer contigs which yield better statistics for profiling. When the analysis does not classify each read instead it merely offers the estimated abundances of the various taxa, a process known as taxonomic profiling.

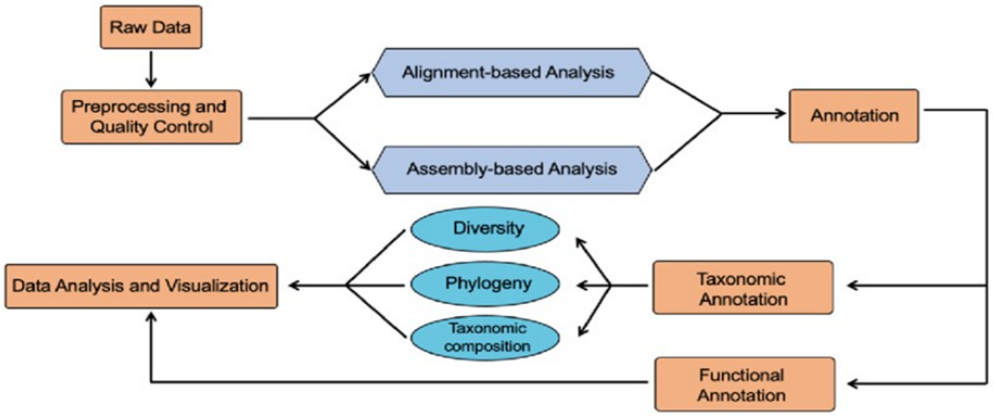


FIGURE 2.5: Metagenomics data analysis techniques.

For quantitative community profiling and for locating closely related organisms in the database, direct taxonomic classification is helpful. Assuming that DNA can be recovered from the target environment, metagenomic shotgun sequencing reduces biases from primer selection and allows the detection of species in all domains of life as compared to marker gene-based community analysis. Using ecological and biogeographic criteria, researchers may measure the species diversity, richness, and homogeneity of microbial communities. While the reference library has many complete genomes of bacteria linked to humans, metagenomics sequencing has identified certain pathogens while comprehensive sequencing has left others undetected [30].

2.9 Biomarkers Discovery

A biological finding that forecasts a significant clinical diagnosis endpoint or intermediate result is called a biomarker. Multiple studies have been carried out

to find T2DM biomarkers. Discovery of biomarkers stands out as one of the most common and effective sources of mapping molecular and genomic data into healthcare practices. The biomarkers discovery employs comparison of healthy vs. case samples and brings the task of detecting novel types and a disease subtype or figuring out a fresh sample's subtype to spotlight. With the development of sequencing technologies, it has now become easier to identify biomarkers, enabling clinical diagnostics and microbiological applications by comparing different microbial communities [20].

2.9.1 Novel Approach of Using Human Gut Microbiome As Non-Invasive Biomarker

Recent years have seen a rise in studies on non-invasive, omics-based T2DM biomarkers which may lead to the discovery of new biomarkers. One of the most effective and widely applicable methods for integrating genomic and molecular data into clinical practice is biomarker discovery. Studying the differences between diseased and healthy tissues has brought attention to the significance of class prediction (figuring out a new sample's subtype) and class discovery (finding new disease subtypes) and the utility of human microbial communities as biomarkers for host variables including lifestyle and illness has been demonstrated by recent metagenomic assays. An important new field for metagenomic biomarker discovery is the human microbiome, which is made up encompasses all the microbiological species connected to human hosts. Variations in the quantity of microorganisms in the skin, mouth, and gut have been linked to a number of disease conditions, including T2DM and obesity. In general, any uncultivated sample's microorganisms or microbial metabolic properties can be identified through the metagenomic analysis of microbial communities. Certain species, clades, functional taxa, or pathways whose relative abundances vary across two or more sample groupings are usually the focus of metagenomic data analyses, and several microbial community characteristics have been proposed as possible indicators of various illness stages. Individual pathogenic organisms, for example, can indicate the presence of disease if they are present in a community, and bacterial vaginosis and Crohn's

disease (CD) have been shown to exhibit increases and decreases in community complexity. Although there is a correlation between illness symptoms and each of these several kinds of microbial biomarkers, there aren't many bioinformatic techniques available to interpret the class comparisons provided by metagenomic data [20].

2.10 Challenges Associated with Identification of Metagenomics Biomarkers

2.10.1 Multi-Dimensional Data

Metagenomic analyses yield huge data that is multi-dimensional and complex to interpret. Thus, extracting biologically relevant attributes and tracing meaningful patterns from such a big data is challenging and can lead to false biomarkers [31].

2.10.2 Technical and Biological Bias

Regardless of the data type or experimental protocols, this high-dimensional natured data experiences several challenges. For instance, the number of biomarkers identified for a given dataset exceeds the total number of samples included in a study. Additionally, metagenomic data inherits technical and biological biases, such as sequencing errors, chimeric reads, complex underlying biology, and genetic variations between replicates [20].

2.10.3 Reproducibility

On the other hand, once the biomarkers are determined, the next step is t to guarantee the repeatability of the findings derived from metagenomic research and is crucial for evaluating the practicality of the biological findings. This is usually accomplished using the available statistical tools.

2.10.4 Generalized Biomarkers

Till now, meta-analysis studies conducted to determine the diagnostic biomarkers across multiple cohorts fail to clarify how well the discovered microbial signatures generalize across different studies of the same disease. Thus, they left it ambiguous if the biomarkers are disease-specific and can diagnose the disease if different test datasets are presented to avoid the over-optimistic evaluation of their prediction accuracy [10].

2.10.5 Limited Resolution of Amplicon Sequencing

Metagenomic analyses carried out for the identification of disease specific biomarkers utilized amplicon sequencing due to its cost-effectiveness. However, recently, meta-analyses based on 16s ribosomal RNA (rRNA) have discovered striking technical variations between studies. The taxonomic biomarkers identified by many studies were shown to either have a low effect size or have unresolved association with the onset and progression of corresponding diseases. On the hand, shotgun metagenomic have been shown to have better taxonomic resolution as well as potential to elucidate the functional biomarkers providing the needed statistical power to map the disease-microbe interactions [10].

2.11 Available Computational Tools for Biomarkers Discovery

Currently available tools that are commonly applied in the identification of biomarkers are illustrated in Table 2.1.

TABLE 2.1: Available computational tools and packages utilized in metagenomic biomarkers discovery

Tool	Description	Ref.
LEfSe	Identifies differentially abundant genes, pathways, and taxa between two or more samples.	[20]

Table 2.1 continued from previous page

Tool	Description	Ref.
RPCA	Randomized sampling-based biomarkers discovery algorithm	[32]
RegLRSD	Models the bacterial abundance as a matrix and provides differentially and non-differentially abundant microbes.	[33]
IMG/M	Microbiome data management system that facilitates biomarkers discovery by providing annotated datasets of bacteria, archaea, and fungi.	[34]
MeAtML	Machine learning based tool for metagenomics-based prediction tasks including mapping of the microbiome-phenotype associations.	[35]
Fizzy	Predicts OTUs and functional biomarkers by implementing feature selection techniques.	[36]
Boruta	R package implementing novel feature subset selection algorithm to selected biologically relevant biomarkers.	[37]

2.12 Taxonomic Biomarkers of T2DM

There is much evidence linking dysbiosis in the gut microbiota to the beginning and development of T2DM. Previously, studies showed few members of the gut microbiota such as Faecalibacterium, Akkermansia and Roseburia to have negative associations with T2DM implying decreased abundance of these micro-organisms increases the susceptibility of the onset of T2DM. These microbes cause the disease onset by increasing insulin resistance, decreasing the gut permeability, and affecting host metabolism and signaling pathways. This suggests that restoring the concentration of these specific microbes to the specific threshold can ameliorate the intensity of T2DM aiding in its management and reversal [34]. In consensus with this study, another study has also found decrease in the abundance of Akkermansia muciniphila causing the onset of diseases. It suggested that the functional changes reflecting response to oxidative stress is also involved in the onset of T2DM. In another study, several taxonomic biomarkers such as an increase abundance of E. coli, and S. salivarius, have been determined for diagnosing T2DM [38]. Additionally, a decrease in Bacteroides vulgatus and Bacteroides

abundant uniforms has been reported [39]. Apart from the taxonomic biomarkers, several research studies have compiled information on T2DM risk biomarkers, both genetic and non-genetic. Additionally, omics-based, non-invasive biomarkers for type 2 diabetes have been the focus of research in recent years, and numerous novel biomarkers have been proposed [40]. Thus, literature review illustrates that, most of the studies conducted to determine the T2DM biomarkers amplicon sequencing based and limits the resolution of taxonomic biomarkers up to genus or specie level. Also, the identified T2DM biomarkers are generalized and do illustrate disease-specific characteristics.

2.13 Read-Based and Assembly-Based Metagenomics

Read-based approaches in metagenomics play a crucial role in accurately identifying the sources of reads, especially in long-read sequencing where high error rates can obscure origins [41]. Long-read sequencing technologies have significantly improved read level accuracy, enabling the recovery of diverse biosynthetic gene clusters (BGCs) from uncultivated bacteria in environmental samples like seawater [42]. Benchmarking studies comparing long-read assemblers like Flye, Raven, and Redbean have shown that increasing read depth benefits assembly quality, with Flye being the most robust and effective assembler for recovering plasmids in metagenomic studies [43]. Additionally, the use of long-read nanopore sequencing for pathogen detection in clinical settings has highlighted the importance of gentle DNA extraction methods to obtain unbiased and integrated DNA, showcasing the superior performance of enzymatic-based methods in pathogen identification and disease diagnosis [44].

Read-based taxonomy, such as sequence alignment and marker genes analysis, plays a crucial role in microbiome research and plant genotyping. Utilizing Illumina paired end reads by joining them instead of merging can enhance taxonomy

annotation accuracy, especially when dealing with sequencing errors, thus maximizing the potential of the data for classification. Marker gene analysis is common for taxonomic profiling in microbiome research, but predicting functional potential from short-read amplicons may lack biological validity due to taxonomic resolution and data composition [45].

There are several approaches for taxonomic profiling of metagenomic data, which can be broadly categorized into three main methods: genome-based, gene-based, and k-mer based [46].

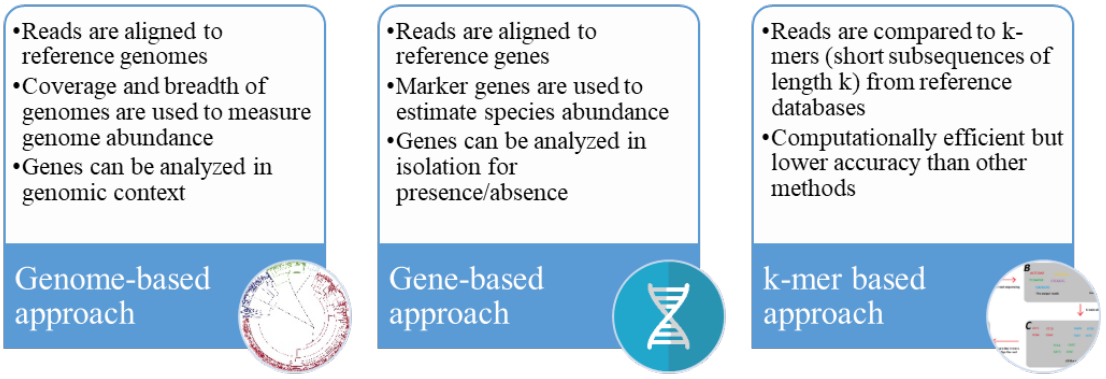


FIGURE 2.6: Approaches typically adopted for the identification of taxonomic biomarkers.

Some specific tools and methods for read-based taxonomy include:

1. **Marker gene based:** Searching for marker genes like 16S rRNA in reads, which is quick but introduces bias. Tools like MetaPhlAn use this approach.

2. **Sequence alignment:** Reads are aligned to reference databases of DNA or protein sequences. DNA-to-DNA comparison is done by tools like Kraken, while DNA-to-protein comparison (more computationally intensive) is done by tools like DIAMOND [46].

3. **ORF-based:** Reads are translated to amino acid sequences and compared to reference databases. This allows taxonomic assignment of reads from different genomic regions [47].

4. **Probabilistic scoring:** Tools like MetaMaps use approximate mapping with probabilistic scoring to estimate sample composition from long reads [48].

Assembly-based taxonomic profiling of metagenomic data involves assembling reads into longer contigs and then binning these contigs into metagenome-assembled genomes (MAGs) based on sequence composition and coverage. This approach provides a more complete picture of microbial communities compared to read-based methods [49].

Microbial profiling, the identification and quantification of microbes in a sample, can be performed using either read-based or assembly-based methods. Both approaches have their advantages and disadvantages, a comparison table summarizing the key aspects of read-based and assembly-based methods for microbial profiling [49, 50].

2.14 Combining Read-based and Assembly-based Taxonomy

Combining read-based and assembly-based methods can provide a more comprehensive and accurate taxonomic profiling of microbial communities. By leveraging the strengths of both approaches, researchers can obtain higher taxonomic resolution, capture genomic context, and identify novel organisms while still maintaining computational efficiency.

Assembly-based methods, such as binning contigs into metagenome-assembled genomes (MAGs), can provide higher taxonomic resolution compared to read-based methods, allowing for more accurate identification of species and strains. By assembling reads into longer contigs, assembly-based approaches capture genomic context, which is crucial for understanding the functional potential and interactions of microbes within a community [51]. These methods, particularly

when combined with long-read sequencing technologies, can facilitate the discovery of novel genomes and microbes that may be missed by read-based approaches. While read-based methods are computationally efficient and can be used for rapid initial profiling of microbial communities, assembly-based methods can be applied to specific regions of interest for deeper analysis, offering a more comprehensive understanding of microbial diversity and function [52].

2.15 Rationale for Integrating both Methods in T2DM Research

T2DM is a multifactorial metabolic disorder characterized by insulin resistance and impaired glucose regulation. Recent research highlights the role of the gut microbiome in T2DM. Combining read-based and assembly-based methods provides a comprehensive view. Read-based techniques rapidly profile the gut microbiome, identifying known pathogens or beneficial bacteria. Assembly-based methods focus on specific genomic regions, revealing novel microbes and functional genes associated with T2DM. These insights may lead to targeted therapies and diagnostic biomarkers [49].

Previous research has combined read-based and assembly-based approaches to study the gut microbiome in T2DM. By leveraging the strengths of both methods, researchers can obtain a more comprehensive understanding of the taxonomic composition and functional potential of the gut microbiome in T2DM patients [53]. Liu et al. used a multi-omics network approach to uncover the gene regulatory network of T2DM, identifying signature genes, biomolecular processes, and pathways, which revealed potential novel strategies for pharmacologic interventions [53]. Xu et al. compared the assembly quality of long-read (PacBio) and short-read (Illumina) sequencing data for soil metagenome exploration and found that combining both methods improved efficiency, yielding higher numbers of genes, functional proteins, and microbial diversity than either method alone [52].

Combining read-based and assembly-based methods for taxonomic profiling of the gut microbiome in T2DM research can provide several benefits, but it also comes with potential pitfalls to consider. By assembling reads into longer contigs and binning them into metagenome-assembled genomes (MAGs), researchers can achieve higher taxonomic resolution compared to read-based methods alone, allowing for more accurate identification of species and strains associated with T2DM.

Combining read-based and assembly-based methods for taxonomic profiling of the gut microbiome in T2DM research can provide several benefits, but it also comes with potential pitfalls to consider. By assembling reads into longer contigs and binning them into metagenome-assembled genomes (MAGs), researchers can achieve higher taxonomic resolution compared to read-based methods alone, allowing for more accurate identification of species and strains associated with T2DM. This assembly process also captures genomic context, crucial for understanding the functional potential of the gut microbiome in T2DM, such as genes involved in glucose metabolism or inflammation. Additionally, assembly-based methods facilitate the discovery of novel genomes and microbial genes linked to T2DM, potentially identifying new therapeutic targets or diagnostic biomarkers. By combining read-based and assembly-based methods, researchers can gain complementary insights, providing a more complete picture of the gut microbiome's composition and functional potential in T2DM, enhancing the understanding of the microbiome's role in the disease's development and progression [54].

Combining read-based and assembly-based methods requires significant computational resources, including high-performance computing and storage capacity, which can limit some research groups. The data analysis is complex and time-consuming, requiring bioinformatics expertise and specialized software tools.

The success of taxonomic profiling depends on the completeness and accuracy of reference databases; incomplete or erroneous databases can lead to misclassification. Interpreting results from these combined methods is also challenging, especially with complex or novel microbial communities, necessitating careful consideration of each method's limitations and the use of appropriate statistical and bioinformatics tools [55].

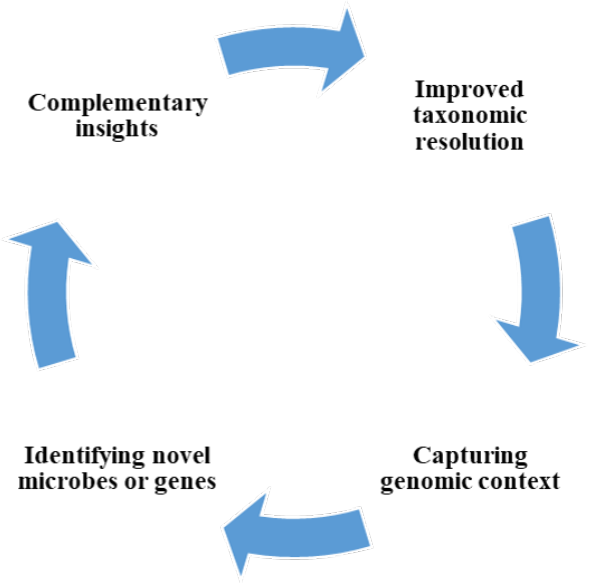


FIGURE 2.7: A typical workflow cycle for metagenomic biomarkers identification.

2.16 Cross-population Studies in Metagenomics

Cross-population studies in metagenomics aim to compare the gut microbiome composition and functional potential across different populations. These studies provide insights into the impact of host genetics, diet, lifestyle, and environmental factors on the gut microbiome. By analyzing data from multiple populations, researchers can identify consistent patterns and variations in the gut microbiome that may be associated with health and disease [56]. A study by Rampelli et al. compared the gut microbiomes of urban and rural populations in Africa and Europe, finding that rural populations had higher microbial diversity, linked to higher fiber intake and lower prevalence of non-communicable diseases. Gupta et al. analyzed gut microbiomes from individuals in the United States, China, and India, identifying population-specific differences in microbial taxa and functional pathways associated with diet and lifestyle. Falony et al. conducted a meta-analysis combining data from multiple studies to create a reference database for the healthy human gut microbiome, revealing that age, country, and body mass index are the main factors influencing gut microbiome composition [29, 57].

Cross-population studies are valuable for identifying a core set of microbes consistently present across different populations, highlighting their role in maintaining a healthy gut microbiome. These studies reveal population-specific differences influenced by diet, lifestyle, or genetic background and explore the impact of environmental factors by comparing populations from different regions [58]. They also help validate associations between gut microbiome composition and health outcomes, strengthening evidence through consistent findings across populations. Metagenomics is effective for examining the genetic diversity of unculturable microbiota, with amplicon metagenomics focusing on microbial diversity and shotgun metagenomics on functional diversity. The combination of metagenomic and next-generation sequencing techniques has also advanced the study of microbial diversity in extreme environments, bypassing the need for traditional culturing methods [59].

Previous studies have compared gut microbiomes across different populations, providing insights into the impact of host genetics, diet, lifestyle, and environmental factors on the gut microbiome composition. Studies have identified a core set of microbes consistently present across different populations, highlighting their importance in maintaining a healthy gut microbiome. Cross-population research has revealed variations in gut microbiome composition influenced by diet, lifestyle, or genetic background [60]. Comparing populations from different regions has elucidated the role of environmental factors in shaping the gut microbiome. Consistent findings across multiple populations validate associations between gut microbiome composition and specific health outcomes. Additionally, studies in wild rodents have shown that the small intestine harbors a unique microbiome compared to the lower gastrointestinal tract (GIT), indicating similar evolutionary pressures on bacteria across species. Significant differences in the relative abundances of genera were found between the upper and lower GIT and in patients with chronic diseases and varying aortic calcification (AoAC) scores [61][62].

Chapter 3

Materials and Methods

This chapter provides a comprehensive overview of the methodology (Figure 3.1) adopted for the metagenomic analysis of human gut microbiome as a tool towards non-invasive biomarkers for diagnosing T2DM.

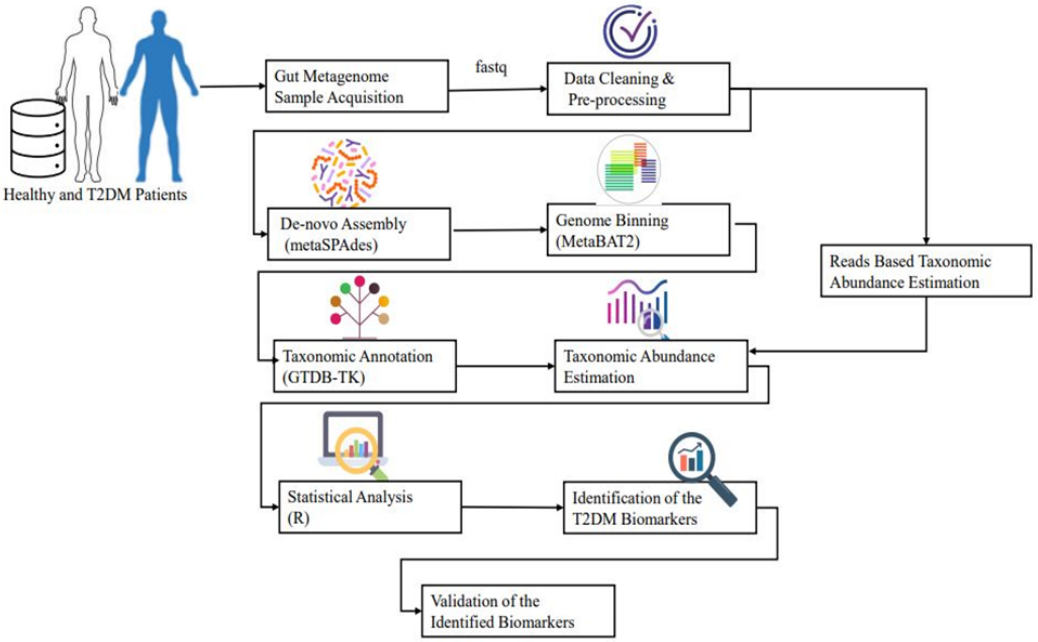


FIGURE 3.1: Methodology for disease-specific biomarkers discovery for T2DM.

3.1 Metagenomics Dataset Collection

This study employs publicly available human gut metagenome samples recruited from T2DM patients and healthy controls. Compressed Fastq files of paired-end raw reads were downloaded from NCBI Sequence Read Archive (SRA) (www.ncbi.nlm.nih.gov/sra). The dataset comprised of 50 healthy and 50 T2DM patient's metagenomes from Chinese and American cohorts available under the BioProject accession number PRJNA422434 and PRJEB11419 respectively.

The selection criteria of the samples include:

1. Availability of the complete clinical information such as age, gender, and BMI.
2. The sequencing data have been previously published and available.
3. The clinical information matches the sequencing data.

3.2 Data Pre-processing

Data preprocessing is the essential step in the data analysis process. It involves cleaning and transforming the raw data into a format that is suitable for further analysis by removing the host reads, adapter contaminants and filtering out the low-quality base sequences [63].

Following the retrieval of gut metagenomes raw FASTQ files of selected datasets were subjected to quality assessment and preprocessing using a combination of FastQC (v0.12.1) (https://github.com/s-andrews/FastQC) and Fastp tools (v0.23.2) [64].

FastQC is used to generate a quality control (QC) report of the raw sequence data produced by high throughput sequencing technology in HTML format. This report offers a brief evaluation of the quality of the reads and designates them as very abnormal, extremely odd, and normal. The evaluation criteria include

read count, GC content, base count, adapter content and contamination, per-base quality matrix, over-represented sequences, duplicates reads and k-mers analysis. The following FastQC command was used in linux environment:

fastqc –threads 48 –outdir ./preQC ./Raw/sample1.fastq.gz ./Raw/sample1_2.fastq.gz

where –threads refer to the number of processors utilized while running the operation and –outdir refers to the location to store the output file. Following the generation of the QC report, raw reads were cleaned and processed using Fastp. Fastp trimmed the adapter sequences, filtered out the contaminants and low-quality reads (Q<20), yielding high-quality data which was used for further downstream analyses.

fastp –in1 ./Raw/Sample_1.fastq.gz –in2 ./Raw/Sample_2.fastq.gz –out1 ./Clean/Sample1_1.fastq.gz –out2 ./Clean/Sample1_2.fastq.gz –detect_adapter_for_pe –disable_quality_filtering –length_required 120

where –in1 and –in2 refers to the paired end input files and –out1 and –out2 refers to the corresponding output files. –detect_adapter_for_pe was used for detecting adapter sequences for the pair end data and –disable quality filtering was used to protect sequences from filtering out because of quality check –length_required referred to the required length (bp) of the reads after preprocessing.

3.3 *De-Novo* Metagenome Assembly

Metagenome assembly is typically de novo and involves direct reconstruction of genome using reads without any prior knowledge enabling the recovery of unculturable microbes [65]. Based on the number of reads and the complexity of the microbial species in the sample, MAGs are assembled from short reads either as a

complete or a draft genome [65]. *De novo* assembly is based on either the overlap-layout-consensus (OLC) and De Bruijn graph (DBG) based approaches where the DBG is the most prevalent method [65].

Hence, this study has also used the DBG based assembler i.e. metaSPAdes (v3.15.3) that uses k-mers for assembling the selected metagenomes. metaSPAdes was opted because of its potential to yield highest N50 contigs and best assembly performance especially in case of complex metagenomes. The assembler was run with default parameters (k-mers) and keeping minimum contig length of 2000 bp (Figure 3.2).

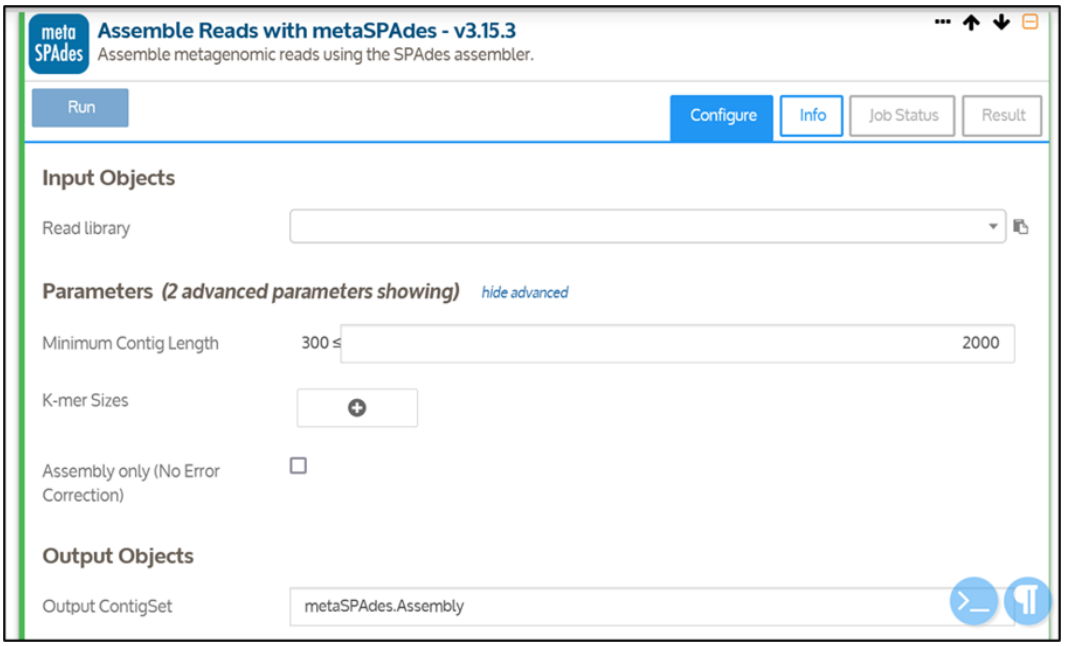


FIGURE 3.2: Screenshot of metaSPAdes tool showing the employed parameters.

3.4 Genome Binning

Genome binning is performed to classify the resulting contigs or scaffolds into separate bins where each bin represents a metagenome assembled genome (MAG). The resulting contigs were binned using MetaBAT2 (v1.7) at default settings (Figure 3.3) and minimum contig length of 2500bp. MetaBAT2 is the most widely used genome binner that reconstructs individual genomes from microbial communities and can handle huge metagenomic datasets.

Genome binning yields bins of varying qualities (such as draft or high-quality), which need to be post-processed (such as refined and dereplicated) before being subjected to downstream analysis.

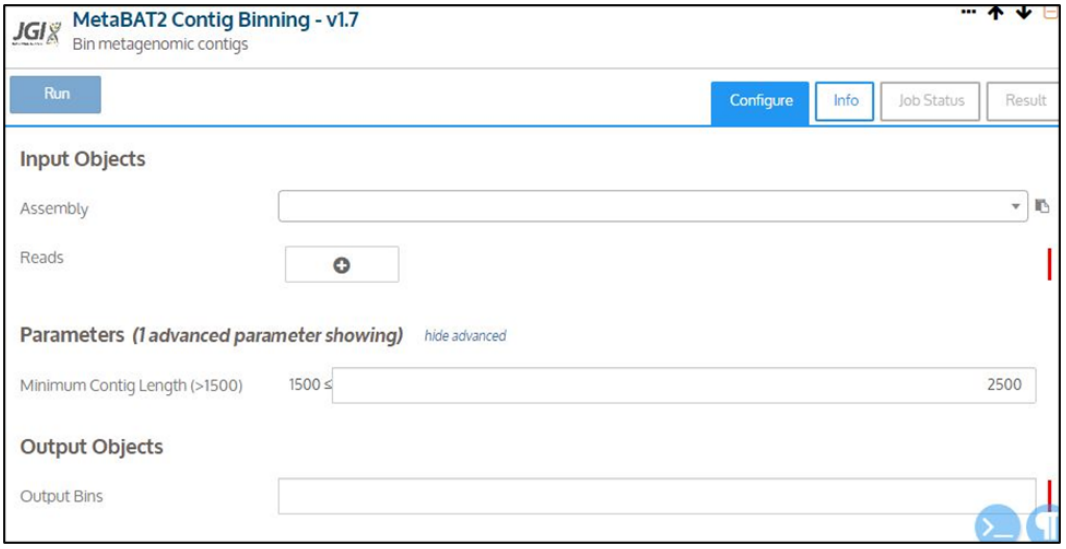


FIGURE 3.3: Screenshot of MetaBAT2 showing the default parameters.

3.5 Taxonomic Annotation and Abundance Estimation

To infer the composition and relative abundances of the microbial communities, present in the metagenomic samples either the assembly-based or reads-based approach is adopted. A read based approach classifies reads and assigns them to different taxonomies while the assembly-based approach annotates the recovered MAGs. However, a read-based method is widely being deployed in biomarker studies as it is faster and accurately estimates the relative abundances of the taxa within metagenomes. Here, an integrative approach i.e. using both read-based and assembly-based method was used to assess the microbial communities along their relative abundances.

For that first, high quality reads were subjected to taxonomic profiling using Kaiju (v1.9.0). Kaiju is the most accurate reads classifier and classifies 10X more reads than other metagenome classification tools. It uses the NCBI taxonomy and a

reference database of protein sequences from viral and microbial genomes and assign reads directly to the taxa. Taxonomic abundances were estimated at genus and specie level in Greedy mode (Figure 3.4). Reference database used was RefSeq Genomes and low abundance filter was set to >=0.05. Taxonomic abundance profiles were generated at both genus and species level.

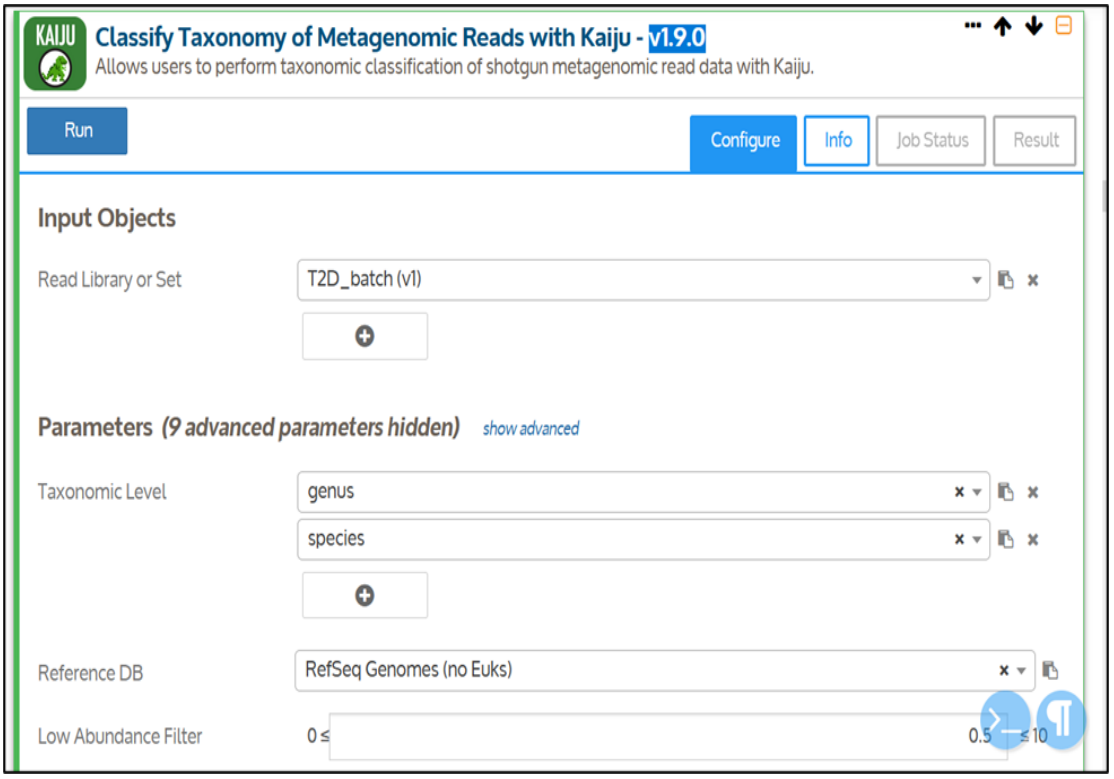


FIGURE 3.4: Screenshot of kaiju tool showing the default parameters.

In the next step, a software toolbox called GTDB-Tk (v2.3.2) was used to annotate the bacterial genomes according to their taxonomy. GTDB-TK was used to annotate the recovered bins at Greedy mode (Figure 3.5).

Recovery threshold value of 0.05 was used to include the low abundance species along with the most prevalent and highly abundant taxa. Later, the results from two approaches were combined to get the final taxonomic profiles of healthy and T2DM patients.

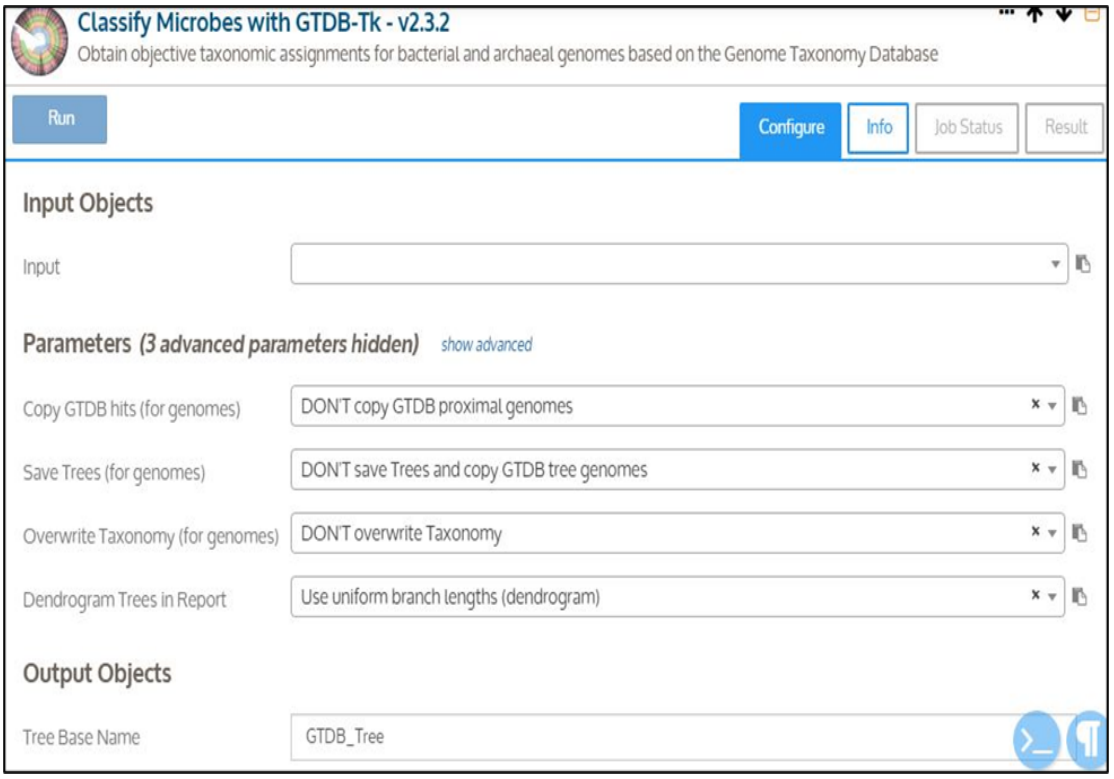


FIGURE 3.5: Screenshot of GTDB-TK showing the default parameters.

3.6 Statistical Analysis and Biomarkers Identification

The Wilcoxon Rank Sum test was performed to record significant differences between the control and T2DM gut metagenome profiles at genus as well as species level in R. Differentially abundant taxa between the control and T2DM metagenomes were selected as a potential microbial biomarker.

3.7 Platform

Quality evaluation and data pre-processing steps were carried out in linux environment using command-line programs. Whereas the downstream analysis from de-novo assembly to MAGs taxonomic annotation was conducted on kbases. Kbase (https://www.kbase.us/) is a web-based, easy to use, platform that integrates a

variety of data analysis tools with an objective to help scientists work together to uncover new information with the purpose of predicting, controlling, and designing the function of microorganisms, plants, and their communities. Additionally, statistical analysis was conducted in R programming languages using R-Studio.

Chapter 4

Results and Discussion

This chapter covers the results achieved and the discussion in context of the aims and objectives of the study. The study aimed at the gut metagenome profiling of human control and T2DM patients to identify the disease biomarkers irrespective of the host genetics. This research employed metagenome samples from the cohorts where the T2DM is the most prevalent disease i.e. Chinese and Americans. By conducting the multi-cohort study, this work tried to focus on the disease specific biomarkers by overlooking the possible genetic and technical variations.

4.1 Data Cleaning and Pre-processing

To conduct an accurate analysis, raw reads were subjected to the data cleaning and pre-processing process.

4.1.1 Evaluating the Quality of the Sequencing Data

Here, first, quality reports of raw sequencing data were obtained using fastQC. The generated reports illustrated that some of the reads of the input samples contained adapter contamination, problematic G+C content as well as per base sequence content. Also, some samples showed reads with phred scores less than

20 that were also required to be removed. Phred scores referred to as q score is a quality measure assigned to each read in the sequencing data and estimates the probability of bases being called out incorrectly. It is presented on a negative scale such that higher q-scores refers to a more accurate base call [66].

Briefly, if the q score is less than 20, it refers to substandard sequencing reads whereas, a q-score greater than 20 means reliable reads. Example QC reports of a few samples are shown in Figure 4.1 – Figure 4.3. These results required us to clean the reads which were achieved using fastP in the next step.

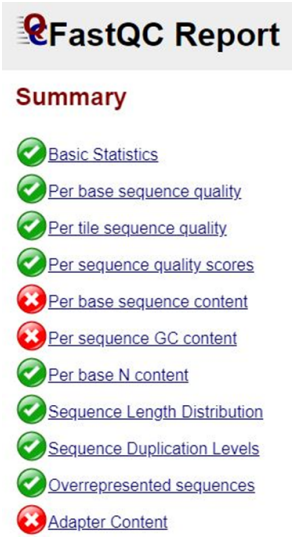


FIGURE 4.1: Summary of FastQC report indicating problematic per-base sequence content, G+C content and presence of adapter contamination.

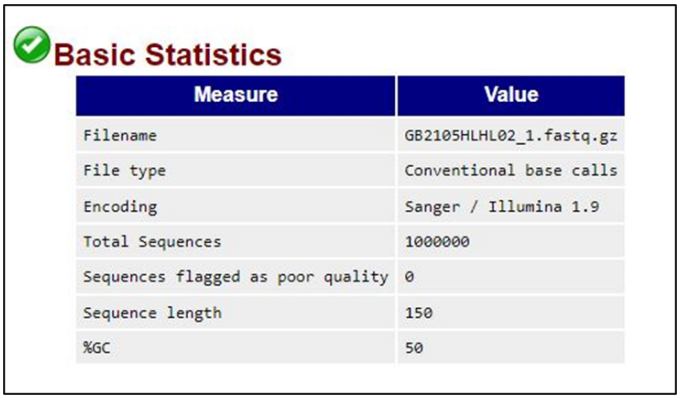


FIGURE 4.2: Basic statistics summary from FastQC report providing information like sample name, total sequences and sequence length of the sample.

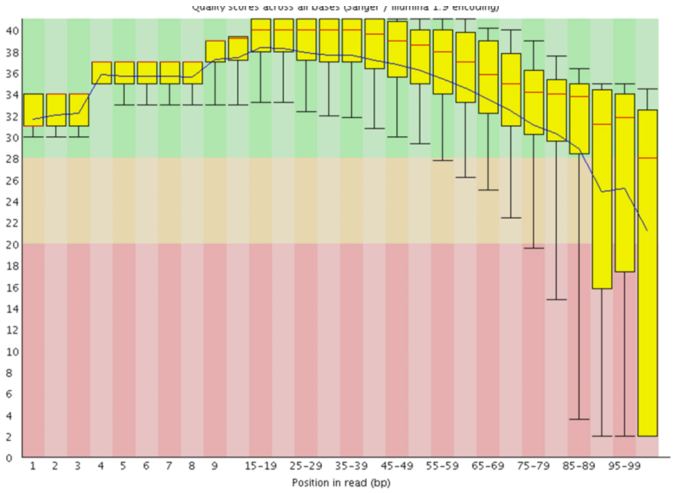


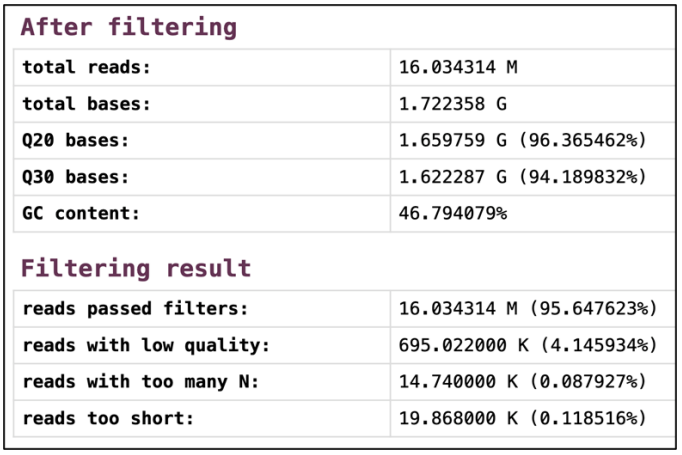
FIGURE 4.3: Quality scores across all bases indicating that quality declines after 74 (bp) in reads.

4.1.2 Cleaning the Raw Metagenomic Sequencing Data

Considering the quality control report generated by FastQC, there was a need to trim the adapter sequences, and low-quality bases from the reads. For this reason, we used fastP which is the most well-known preprocessing tool. This step was again conducted in linux environment and produced reads with mean length of 107 bp and 96% of bases having q-scores >20. fastP when processing the reads, generates a quality report alongside which indicates the quality metric of reads before and after processing. The fastP report of a few samples before and after processing is shown in Figure 4.4 and Figure 4.5. After conducting the data processing, reads were further referred to as clean reads.

fastp version:	0.17.0 (https://github.com/OpenGene/fastp)
sequencing:	paired end (151 cycles + 151 cycles)
mean length before filtering:	108bp, 108bp
mean length after filtering:	107bp, 107bp
duplication rate:	30.641418%
Insert size peak:	95
Before filtering	
total reads:	16.763944 M
total bases:	1.818801 G
Q20 bases:	1.716550 G (94.378124%)
Q30 bases:	1.672955 G (91.981195%)
GC content:	47.006320%

FIGURE 4.4: fastP report showing the status of sequencing data before filtering.



After filtering	
total reads:	16.034314 M
total bases:	1.722358 G
Q20 bases:	1.659759 G (96.365462%)
Q30 bases:	1.622287 G (94.189832%)
GC content:	46.794079%
Filtering result	
reads passed filters:	16.034314 M (95.647623%)
reads with low quality:	695.022000 K (4.145934%)
reads with too many N:	14.740000 K (0.087927%)
reads too short:	19.868000 K (0.118516%)

FIGURE 4.5: fastP report showing the status of sequencing data after filtering.

4.2 De novo Assembly and Recovery of MAGs

De-novo assembly of human gut metagenome samples performed using metaSPAdes had a mean assembly length of 78477.50 ± 29765 Kbp for healthy controls and 79533± 2500 Kbp for T2DM. The increased assembly size of T2DM here corresponds to increased diversity of gut microbiome in diabetics. However, this difference is not significant as determined by Wilcoxon rank sum test (P >0.05) Test.

Subjecting these assemblies to the process of genome binning yielded 120 and 170 MAGs for healthy and T2DM patients. These MAGs were subjected to quality assessment using CheckM and classification into high quality (HQ), medium quality (MQ) and low-quality (LQ) MAGs using MIMAG criteria. Results are presented in Table 4.1.

TABLE 4.1: Quality of MAGs recovered from human gut metagenome of healthy and T2DM patients.

Condition	**Healthy**	**T2DM**
MAG Count	1591	1721
HQ%	35.4	35.7
MQ%	45.6	55.0
LQ%	19.0	9.3

4.3 Taxonomic Annotation and Profiling

Natural microbial communities are composed of a variety of different microorganisms that can be directly determined and characterized through their DNA with shotgun metagenomic sequencing technique [67]. To determine the overall composition and relative abundance of microorganisms present in a metagenomic sample, taxonomic annotation is performed generating taxonomic profiles. The resulting taxonomic profiles then illustrate the proportion of DNA contributed by a specific member of a community to the metagenome [68]. When it comes to the taxonomic profiling, two approaches can be opted:

1. ***Read-based approach*** refers to the direct classification of the reads by aligning them against the reference databases.

2. ***Assembly-based approach*** refers to the alignment of reads into contigs which are then subjected to genome binning to recover high-quality draft metagenomes followed by taxonomic annotation using different tools.

Both the above approaches have their pros and cons and can be preferred depending upon the objectives of the study. If the aim is to just classify the reads and know which microbial taxa are present in the samples, reads-based technique is appropriate whereas if the objective is to recover microbial genomes and study their functional potential, assembly-based method is more suitable [69]. The taxonomic profiles depend upon the taxonomic annotation for their accuracy. Though, for most of the datasets, MAGs and contigs can be annotated more reliably than the individual reads, it is noteworthy that all reads are not assembled into contigs. Moreover, all contigs cannot be binned into MAGs [70]. On the other hand, reads can be assigned to different taxonomies but cannot recover the microbial genomes. Therefore, to obtain the accurate taxonomic profiles, an integrated approach utilizing both reads based classification and assembly-based metagenomes annotation was opted here.

4.3.1 Healthy Gut Metagenomes

Read-based taxonomic profiling of 50 healthy gut metagenomes resulted in 142 genera (Figure 4.6) and 234 species (Figure 4.7).

4.3.1.1 Genus Level Profiles

The taxonomic profiles of the control group at genus levels show lesser diversity in microbial communities as compared to the T2DM group. Analysis of the resultant taxonomic profiles revealed the highest proportions of Bacteroides and Phocaeicola (27.2 % and 17.2% respectively), constituting the healthy gut metagenomes. Among 142 identified genera, 10 were present in high abundance >1% and rest 132 were present at the relative abundance of <1%. *Acinetobacter*, *Desulfitobacterium*, *Caproiciproducens*, and *Niabella* are among the low abundance genera and are present at 0.001%. Top 20 most prevalent and low abundance taxa are shown in Table 4.2.

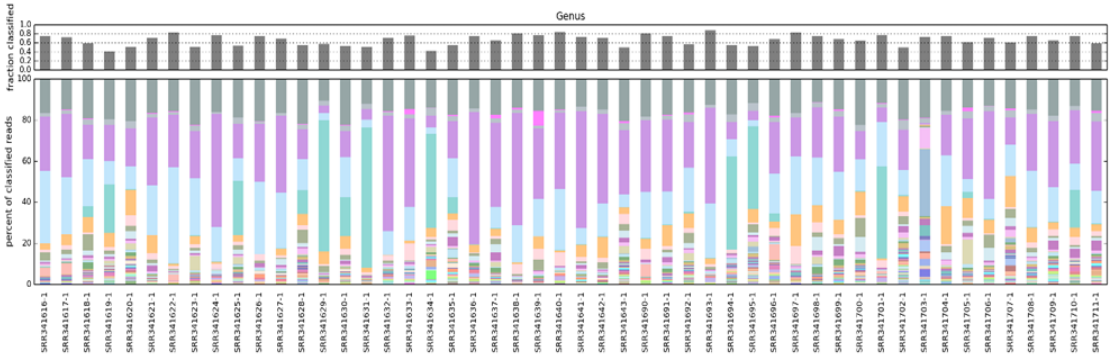


FIGURE 4.6: Genus level diversity in taxonomic profiles of healthy control group.

Bacteroides is the highly abundant genus found in human intestine and mainly plays its role in the polysaccharide's degradation via its polysaccharide utilization loci which contains closely related gene clusters associated with polysaccharide catabolism [71]. *Phocaeicola* being the second most abundant genera is involved in various important functions such as utilizing the carbohydrates, synthesis of vitamins and hormones. Most importantly, this genus has an ameliorating effect

on inflammation, and reduces atherosclerosis by modulating gut microbiota and regulating the cytokines level [72]. *Actinetobacter* is a gram-negative, opportunistic genus with relatively unknown functions in the human gut microbiome and serves as a reservoir for infections [73]. *Desulfitobacterium* normally detected in low abundance in fecal samples is involved in detoxification of toxins in the gut. Its presence is generally associated with a higher dietary intake of fiber and whole grains [74]. *Caproiciproducens* is thought to play a role in the sugar fermentation and other complex carbohydrates and convert them down into simpler molecules so that they can be absorbed by the human gut [75].

TABLE 4.2: Top few genera constituting the taxonomic profiles of healthy control group.

S. No.	Taxon ID	Genus Name	Relative Abundance (%)
1	816	*Bacteroides*	27.2882947
2	909656	*Phocaeicola*	17.2865808
3	838	*Prevotella*	8.70408576
4	216851	*Faecalibacterium*	5.44614692
5	375288	*Parabacteroides*	2.71177602
6	841	*Roseburia*	2.25197246
7	28050	*Lachnospira*	1.64357628
8	239759	*Alistipes*	1.43738564
9	572511	*Blautia*	1.1865162
10	1730	*Eubacterium*	1.10272226
11	1301	*Streptococcus*	0.87596334
12	33024	*Phascolarctobacterium*	0.79642232
13	1263	*Ruminococcus*	0.65606538
14	158846	*Megamonas*	0.61322704
15	816	*Bacteroides*	27.2882947
16	83614	*Luteimonas*	0.00176922
17	1866885	*Mycolicibacterium*	0.00163328
18	1348611	*Vallitalea*	0.00147058
19	83618	*Pseudoxanthomonas*	0.00146712
20	1849828	*Paeniclostridium*	0.00135514
21	253238	*Ethanoligenens*	0.00135284
22	1839	*Nocardioides*	0.00127074
23	59732	*Chryseobacterium*	0.00123864
24	133925	*Olsenella*	0.0012074

Table 4.2 continued from previous page

S. No.	Taxon ID	Genus Name	Relative Abundance (%)
25	1508657	*Ruminiclostridium*	0.00117738
26	2576755	*Caproicibacter*	0.00117046
27	141948	*Thermomonas*	0.00113668
28	84567	*Pedobacter*	0.0011348
29	1378	*Gemella*	0.00113206
30	28453	*Sphingobacterium*	0.00110866

4.3.1.2 Species Level Profiles

At species level, *Prevotella copri, Faecalibacterium prausnitzii, Phocaeicola vulgatus, Phocaeicola dorei, Bacteroides uniformis, Bacteroides thetaiotaomicron, Bacteroides stercoris, Lachnospira eligens,* and *Bacteroides fragilis* were found in relatively high abundance i.e. 6.23%-1.43%. Whereas *Caproiciproducens sp.* NJN-50, *Streptococcus australis*, and *Streptococcus gallolyticus* were found in low abundance i.e. 0.001% among others. A few of the top prevalent and low abundance species are listed in Table 4.3.

Prevotella copri is found to be prevalent in non-western populations such as Chinese, due to high fiber and low-fat diets. It is linked with maintaining the health status of individuals such as visceral fat reduction and improvements in glucose metabolism [76]. *Faecalibacterium prausnitzii* (*F. prausnitzii*) is another abundant bacterial species (mean 5%) and is a major player of gut microbial communities. Changes in the richness and abundance of this species are associated with many metabolic disorders including T2DM [77]. *P. vulgatus* and *P. dorei* are involved in carbohydrate metabolism. P. vulgatus is involved in the production of short chain fatty acids (SCFA) whereas, *P. dolrei* is involved in the reduction of cholesterol, thus maintaining the health status of an individual [72]. *Streptococcus australis* is involved in tumor progression and thus present in very little abundance (0.001%) in healthy humans [78]. *Lachnospiraceae eligens* is an active player of SCFA production such as butyrate and acetate [79]. *Streptococcus gallolyticus* being a member of the healthy human gut microbiome is involved in carbohydrate fermentation and

production of SCFA. A higher relative abundance of *S. gallolyticus* is associated with colorectal cancer [80].

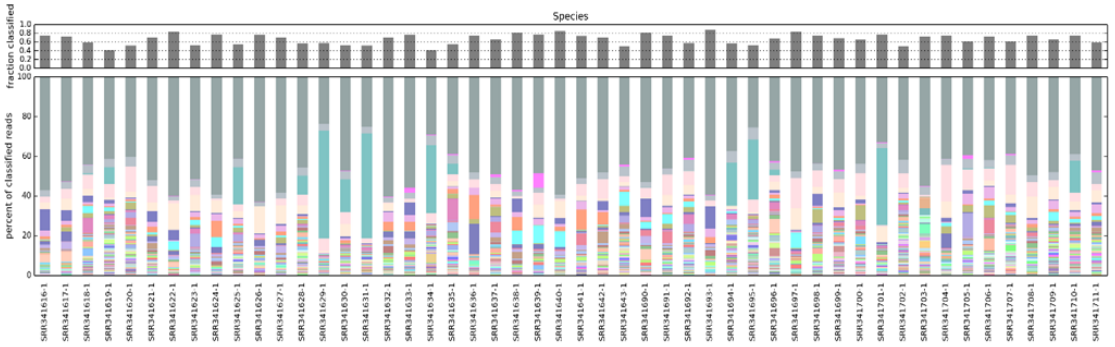


FIGURE 4.7: Species level diversity in taxonomic profiles of healthy control group.

TABLE 4.3: Top few species constituting the taxonomic profiles of healthy control group

S. No.	Taxon ID	Genus Name	Relative Abundance (%)
1	165179	*Prevotella copri*	6.23044204
2	853	*Faecalibacterium prausnitzii*	5.44614692
3	821	*Phocaeicola vulgatus*	4.32878156
4	357276	*Phocaeicola dorei*	2.11174674
5	820	*Bacteroides uniformis*	2.04683496
6	818	*Bacteroides thetaiotaomicron*	1.89369298
7	46506	*Bacteroides stercoris*	1.88684012
8	39485	*Lachnospira eligens*	1.64357628
9	817	*Bacteroides fragilis*	1.43183882
10	387090	*Phocaeicola coprophilus*	0.99566338
11	2841528	*Eubacterium sp. MSJ-33*	0.97722468
12	28111	*Bacteroides eggerthii*	0.96288248
13	28116	*Bacteroides ovatus*	0.90016578
14	246787	*Bacteroides cellulosilyticus*	0.8132772
15	165179	*Prevotella copri*	6.23044204
16	253239	*Ethanoligenens harbinense*	0.00135284
17	417368	*Enterococcus thailandicus*	0.0013522
18	735	*Haemophilus parahaemolyticus*	0.00128522
19	28037	*Streptococcus mitis*	0.00125024
20	2584944	*Sutterella faecalis*	0.00121894
21	2666138	*Caproicibacterium lactatifermentans*	0.00120278

Table 4.3 continued from previous page

S. No.	Taxon ID	Genus Name	Relative Abundance (%)
22	2576756	*Caproicibacter fermentans*	0.00117046
23	1812935	*Enterobacter roggenkampii*	0.00114666
24	43675	*Rothia mucilaginosa*	0.00114296
25	2144175	*Megasphaera stantonii*	0.00108486
26	1156431	*Streptococcus ilei*	0.0010512
27	2086585	*Christensenella sp. Marseille-P3954*	0.0010481
28	876	*Desulfovibrio desulfuricans*	0.0010465
29	28112	*Tannerella forsythia*	0.00103474
30	623	*Shigella flexneri*	0.00103446

4.3.2 T2DM Gut Metagenomes

In the gut metagenome profiles of T2DM patients, 135 genera (Figure 4.8) and 274 species were identified (Figure 4.9). As compared to the control group, diabetic patients gut microbiome was found to be less diverse in terms of genus and more diverse in terms of of species. This dicrepancy can potentially be due to the loss of beneficial genera and acquisition of pathogenic species or opportunistic pathogens involved in the onset of the T2DM.

4.3.2.1 Genus Level Profiles

At genus level, a relative higher proporotion (26% - 0.89%) of *Bacteroides, Phocaeicola, Prevotella, Faecalibacterium, Bifidobacterium, Lachnospira, Alistipes, Blautia, Phascolarctobacterium, Roseburia* and *Mediterraneibacter* was observed. Whereas, *vibrio, Liquorilactobacillus, Pseudoalteromonas, Crassaminicella, Sphaerochaeta, Aeromonas, Ndongobacter, Alkaliphilus,* and *Mageeibacillus* was found to be present at low abundance i.e. 0.001%. A few top high abundance and low abundance genera are presented in Table 4.4.

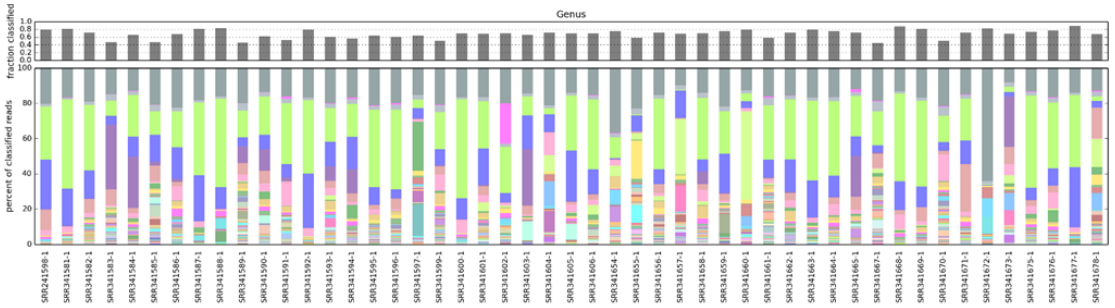


FIGURE 4.8: Genus level diversity in gut metagenome taxonomic profiles of T2DM patients.

A slightly lesser proportion of Bacteroides may indicate diabetes as higher abundance of this genus is positively correlated with improved insulin resisitance in mice model. But this genus has contradictory asscouation with T2DM. Lower proportion of *Ruminococcus, Fusobacterium,* and *Blautia* as compared to the healthy controls found in T2DM gut metagenome profiles are positively associated with T2DM [81]. Genus *Faecalibacterium* is also shown to be reduced in diabetic patients in response to the various antidiabetic medications [82]. Also, a higher abundance of Bifidobacterium as compared to the control illustrates the protective effect in response to the immune system activation [83].

TABLE 4.4: Top few genera constituting the gut metagenome taxonomic profiles of T2DM patients.

S. No.	Taxon ID	Genus Name	Relative Abundance (%)
1	816	*Bacteroides*	26.90512
2	909656	*Phocaeicola*	12.73965
3	838	*Prevotella*	4.379075
4	216851	*Faecalibacterium*	4.183828
5	375288	*Parabacteroides*	3.926047
6	1678	*Bifidobacterium*	3.008247
7	28050	*Lachnospira*	1.800356
8	239759	*Alistipes*	1.759672
9	572511	*Blautia*	1.608178
10	33024	*Phascolarctobacterium*	1.004407
11	841	*Roseburia*	0.923332
12	2316020	*Mediterraneibacter*	0.895504
13	1578	*Lactobacillus*	0.80805

Table 4.4 continued from previous page

S. No.	Taxon ID	Genus Name	Relative Abundance (%)
14	561	*Escherichia*	0.777635
15	906	*Megasphaera*	0.687797
16	158846	*Megamonas*	0.687006
17	86331	*Mogibacterium*	0.002102
18	2767353	*Lancefieldella*	0.002011
19	33870	*Coriobacterium*	0.001994
20	2759736	*Lacticaseibacillus*	0.001689
21	543311	*Parvimonas*	0.001657
22	2743582	*Peptacetobacter*	0.001499
23	662	*Vibrio*	0.001141
24	2767888	*Liquorilactobacillus*	0.001125
25	53246	*Pseudoalteromonas*	0.001115
26	1848399	*Crassaminicella*	0.001075
27	399320	*Sphaerochaeta*	0.001072
28	642	*Aeromonas*	0.001067
29	1930845	*Ndongobacter*	0.001066
30	114627	*Alkaliphilus*	0.001034

4.3.3 Species Level Profiles

The taxonomic profiles of diabetic patients at species level contained more number of species as compared to the control group. The microbial communities constituting the major proportion of the taxonomic profiles were found to be *Faecalibacterium prausnitzii, Prevotella copri, Phocaeicola vulgatus, Bacteroides stercoris, Lachnospira eligens and Phocaeicola dorei.*

These species abundance ranged from 4.1% to 1.5% in the metagenome samples. A few of the species found in the diabetics' gut microbiome profiles are presented in Table 4.5 along with their relative abundance values.

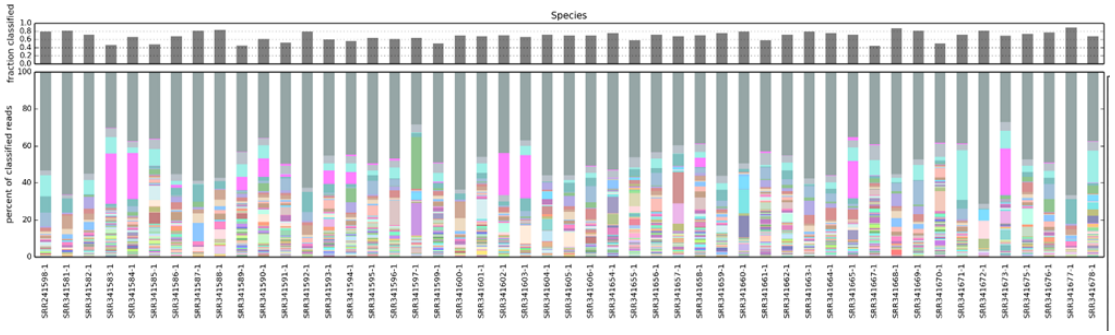


FIGURE 4.9: Species level diversity in gut metagenome taxonomic profiles of T2DM patients.

When compared with healthy controls, increased relative abundance (≈2x of control) of R*uminococcus gnavus, Ruminococcus bicirculans, Ruminococcus bovis,* and *Ruminococcus albus* was determined. Whereas the lower abundance of *Faecalibacterium prausnitzii, Subdoligranulum, Roseburia intestinalis, Roseburia inulinivorans, Ruminococcus, Eubacterium rectale,* and *Akkermansia muciniphila* was identified in the diabetics' metagenomes. These finding are inline with the previously published studies that shows positive association of Ruminococcus and negative association of *Faecalibacterium, Subdoligranulum, Roseburia, Ruminococcus, Eubacterium* and *Akkermansia* at with the onset of the disease[84].

TABLE 4.5: Top few species constituting the gut metagenome taxonomic profiles of T2DM patients

S. No.	**Taxon ID**	**Genus Name**	**Relative Abundance (%)**
1	853	*Faecalibacterium prausnitzii*	4.183827531
2	165179	*Prevotella copri*	3.287139143
3	821	*Phocaeicola vulgatus*	3.198364878
4	46506	*Bacteroides stercoris*	2.37002349
5	39485	*Lachnospira eligens*	1.800356388
6	357276	*Phocaeicola dorei*	1.557138551
7	820	*Bacteroides uniformis*	1.538592408
8	28116	*Bacteroides ovatus*	1.309364184
9	818	*Bacteroides thetaiotaomicron*	1.240282184
10	817	*Bacteroides fragilis*	1.235497
11	47678	*Bacteroides caccae*	1.150367306
12	33025	*Phascolarctobacterium faecium*	1.004407429
13	33038	*Ruminococcus gnavus*	0.895503551

Table 4.5 continued from previous page

S. No.	Taxon ID	Genus Name	Relative Abundance (%)
14	28111	*Bacteroides eggerthii*	0.788160755
15	371601	*Bacteroides xylanisolvens*	0.744357571
16	83683	*Lactobacillus amylolyticus*	0.001208102
17	1302	*Streptococcus gordonii*	0.001173122
18	1338	*Streptococcus intermedius*	0.001165327
19	1679444	*Akkermansia glycaniphila*	0.001138714
20	1515	*Acetivibrio thermocellus*	0.001136143
21	1305	*Streptococcus sanguinis*	0.001133735
22	1358	*Lactococcus lactis*	0.001098653
23	1463165	*Klebsiella quasipneumoniae*	0.001093102
24	94869	*Clostridium gasigenes*	0.001075327
25	1871025	*Ndongobacter massiliensis*	0.001065551
26	288965	*Acetivibrio clariflavus*	0.001059204
27	433296	*Aminipila butyrica*	0.001053735
28	35787	*Limosilactobacillus pontis*	0.001049857
29	1499973	*Escherichia marmotae*	0.001037837
30	884684	*Mageeibacillus indolicus*	0.00102798

4.4 Microbial Biomarkers of T2DM

Gut microbiota modulates inflammation, interacts with the dietary constituents and influences gut barrier, glucose and lipid metabolism, insulin sensitivity, and maintains homeostasis in the human body. For the identification of biomarkers, the analysis of integrative taxonomic profiles shows that a greater number of taxa are recovered via a reads-based approach as compared to the assembly-based approach (Table 4.8). The differential analysis of control and T2DM gut metagenome taxonomic profiles at species and genus levels shows the presence of thirty-two genera exclusively in the control group and twenty-five genera being present only in the diabetic samples (Figure 4.10 and Table 4.6). On the other hand, at species level, thirty-seven were found exclusive to the control group, whereas seventy-six species were determined in T2DM samples (Figure 4.11 and Table 4.7). The presence of these various genera and species exclusively in control

or T2DM gut metagenomes represents the potential of these microbial communities to be utilized as non-invasive biomarkers for the diagnosis of T2DM.

The association of gut microbiota with T2DM has served as a motivation for exponential advancements in scientific studies. With an aim to alleviate the T2DM complications, this study aimed at the identification of the microbial signatures that can be used as non-invasive biomarkers for the diagnosis of the disease. By employing the differential analysis technique of gut metagenome profiles of healthy and control patients, potential T2DM biomarkers were identified. These biomarkers can be further validated in vitro and in vivo.

In the Chinese cohort, the Bacteroidetes/Firmicutes ratio in T2DM was positively and significantly correlated with plasma glucose concentration. Still, it appeared independent of body weight, confirming its association with reduced glucose tolerance.

The identified species biomarkers include *Bacteroides dorei, Bacteroides grailis, Bacteroides stercoris, Bacteroides ovatus, Bacteroides thetaiotaomicron, Bacteroides uniformis, Bacteroides vulgatus, Bacteroides xylanisolvens, Escherichia lenta, Escherichia coli, Faecalibacterium prausnitzii, Lachnospiraceae, Parabacteroides distasonis, Ruminococcus torques* and *Subdoligranulum*.

Bacteroides, being the most abundant genus, is an important indicator of T2DM with its lower abundance. This suggests *Bacteroides* can manage and reverse T2DM by improving glucose intolerance and insulin resistance. Specifically, *B. vulgatus* and *B. dorei* are involved in the upregulation of the expression of tight junction genes in the colon area. A lower abundance of these species leads to decreased gut epithelial permeability along with the decreased production of lipopolysaccharides and promotes endotoxemia [85]. *B. thetaiotaomicron* decreases the production of Th1, Th2, and Th17 cytokines. Summarizing the results, it is concluded that *Bacteroides* species are an important players of glucose metabolism in human and reduction in the abundance of this genus is negatively associated with the onset of T2DM [86].

Bacteroides, being the most abundant genus, is an important indicator of T2DM with its lower abundance. This suggests *Bacteroides* can manage and reverse T2DM by improving glucose intolerance and insulin resistance. Specifically, *B. vulgatus* and *B. dorei* are involved in the upregulation of the expression of tight junction genes in the colon area. A lower abundance of these species leads to decreased gut epithelial permeability along with the decreased production of lipopolysaccharides and promotes endotoxemia [85]. *B. thetaiotaomicron* decreases the production of Th1, Th2, and Th17 cytokines. Summarizing the results, it is concluded that *Bacteroides* species are an important players of glucose metabolism in human and reduction in the abundance of this genus is negatively associated with the onset of T2DM.

The next important microbial community found at lower abundance in T2DM patients is genus *Faecalibacterium*. *F.prausnitzii* is involved in secreting anti-inflammatory peptides and hence negatively associated with T2DM. It is also reported to be a common constituent of probiotics for colitis and thus can be considered as a potential taxonomic biomarker of T2DM [87].

Lower numbers of *P.distasonis* were also evident in T2DM guts. *P.distansonis* is involved in the production of succinate and secondary bile acids, thus preventing obesity and metabolic dysfunctions. Thus, indirectly involved in the onset of T2DM and is suggested as a potential diagnostic biomarker [88].

Lachnospiraceae being the core constituent of the gut is involved in the production of SCFA. It increased abundance negatively affects glucose metabolism and is thus associated with T2DM [88]. *Bifidobacetrium* is the beneficial genera and is plays protective role against disease. Its lower abundance was found in gut metagenome profiles of T2DM patients. *Bifidobacterium* can be utilized in combination as a diagnostic biomarker, as it is more appropriately classified as a marker of general dysbiosis [89]. Rest, opportunistic pathogens like *Clostridium hatheway, Bacteroides caccae, E. coli, and E.lenta* are found in T2DM profiles and are known to increase diabetes [88].

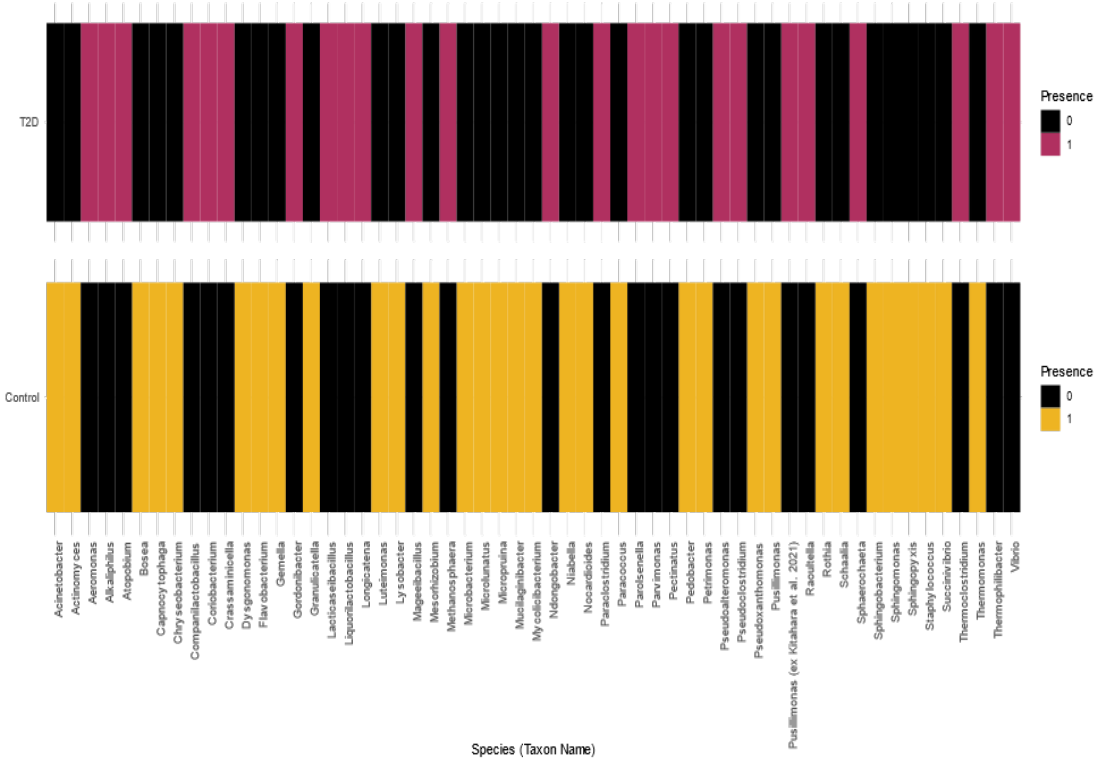


FIGURE 4.10: Genus biomarkers found exclusively in control and T2DM gut metagenomes

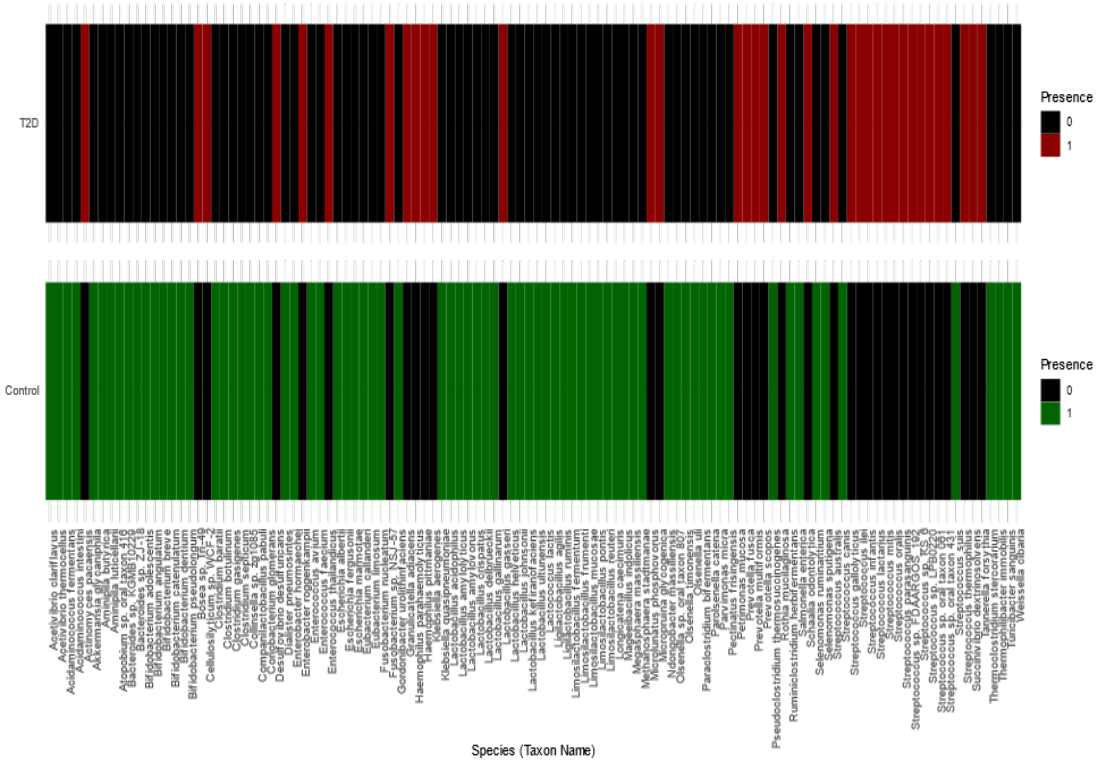


FIGURE 4.11: Species biomarkers found exclusively in control and T2DM gut metagenomes

TABLE 4.6: Genera found exclusively in control and T2DM gut metagenomes.

T2DM		**Healthy**	
Taxon ID	**Genus**	***Taxon ID***	**Genus**
642	*Aeromonas*	68	*Lysobacter*
662	*Vibrio*	237	*Flavobacterium*
864	*Pectinatus*	265	*Paracoccus*
1380	*Atopobium*	469	*Acinetobacter*
2316	*Methanosphaera*	1016	*Capnocytophaga*
33870	*Coriobacterium*	1279	*Staphylococcus*
53246	*Pseudoalteromonas*	1378	*Gemella*
114627	*Alkaliphilus*	1654	*Actinomyces*
160674	*Raoultella*	1839	*Nocardioides*
399320	*Sphaerochaeta*	13687	*Sphingomonas*
543311	*Parvimonas*	28453	*Sphingobacterium*
644652	*Gordonibacter*	29404	*Microlunatus*
1637257	*Mageeibacillus*	32207	*Rothia*
1848399	*Crassaminicella*	33882	*Microbacterium*
1849822	*Paraclostridium*	59732	*Chryseobacterium*
1918536	*Longicatena*	68287	*Mesorhizobium*
1930845	*Ndongobacter*	83614	*Luteimonas*
2082587	*Parolsenella*	83618	*Pseudoxanthomonas*
2304691	*Thermoclostridium*	83770	*Succinivibrio*
2304693	*Pseudoclostridium*	84567	*Pedobacter*
2759736	*Lacticaseibacillus*	85413	*Bosea*
2767879	*Companilactobacillus*	116071	*Micropruina*
2767888	*Liquorilactobacillus*	117563	*Granulicatella*
2847307	*Thermophilibacter*	141948	*Thermomonas*
2892397	*Pusillimonas*		

TABLE 4.7: Species found exclusively in control and T2DM gut metagenomes

Healthy		**T2DM**	
Taxon ID	**Genus**	***Taxon ID***	**Genus**
83771	*Succinivibrio dextrinosolvens*	1685	*Bifidobacterium breve*
548	*Klebsiella aerogenes*	1623	*Ligilactobacillus ruminis*
1318	*Streptococcus parasanguinis*	1604	*Lactobacillus amylovorus*
2610896	*Streptococcus sp. LPB0220*	97478	*Limosilactobacillus mucosae*

Table 4.7 continued from previous page

Healthy		**T2DM**	
Taxon ID	**Genus**	***Taxon ID***	**Genus**
29405	*Microlunatus phosphovorus*	187327	*Acidaminococcus intestine*
1660	*Schaalia odontolytica*	47770	*Lactobacillus crispatus*
589436	*Prevotella fusca*	1680	*Bifidobacterium adolescentis*
282402	*Prevotella multiformis*	1232428	*Megasphaera massiliensis*
1303	*Streptococcus oralis*	33959	*Lactobacillus johnsonii*
75385	*Micropruina glycogenica*	1587	*Lactobacillus helveticus*
1642646	*Petrimonas mucosa*	905	*Acidaminococcus fermentans*
2497860	*Cellulosilyticum sp. WCF-2*	1352	*Enterococcus faecium*
1308	*Streptococcus thermophilus*	1686	*Bifidobacterium catenulatum*
1839799	*Streptococcus sp. FDAARGOS_192*	1796635	*Longicatena caecimuris*
684066	*Streptococcus lactarius*	1805478	*Olsenella timonensis*
1852377	*Actinomyces pacaensis*	2317	*Methanosphaera stadtmanae*
589437	*Prevotella scopos*	1736	*Eubacterium limosum*
249188	*Haemophilus pittmaniae*	154288	*Turicibacter sanguinis*
2764326	*Fusobacterium sp. NSJ-57*	53442	*Eubacterium callanderi*
712623	*Streptococcus sp. oral taxon 061*	1584	*Lactobacillus delbrueckii*
1596	*Lactobacillus gasseri*	1579	*Lactobacillus acidophilus*
1867715	*Bosea sp. Tri-49*	1335613	*Gordonibacter urolithinfaciens*
68892	*Streptococcus infantis*	971	*Selenomonas ruminantium*
712633	*Streptococcus sp. oral taxon 431*	137591	*Weissella cibaria*
2598457	*Streptococcus sp. KS 6*	*28901*	Salmonella enterica
46124	*Granulicatella adiacens*	2779519	*Thermophilibacter immobilis*
417368	*Enterococcus thailandicus*	1598	*Limosilactobacillus reuteri*
735	*Haemophilus parahaemolyticus*	33945	*Enterococcus avium*
28037	*Streptococcus mitis*	2714036	*Companilactobacillus pabuli*
1812935	*Enterobacter roggenkampii*	227945	*Lactobacillus ultunensis*
43675	*Rothia mucilaginosa*	52242	*Lactobacillus gallinarum*
1156431	*Streptococcus ilei*	1613	*Limosilactobacillus fermentum*
876	*Desulfovibrio desulfuricans*	1307	*Streptococcus suis*
28112	*Tannerella forsythia*	2884447	*Bacteroides sp. KGMB10229*

Table 4.7 continued from previous page

Healthy		**T2DM**	
Taxon ID	**Genus**	***Taxon ID***	**Genus**
113107	*Streptococcus australis*	2507160	*Aminipila luticellarii*
315405	*Streptococcus gallolyticus*	1490	*Paraclostridium bifermentans*

TABLE 4.8: Biomarkers identified by integrative (read based and assembly-based) approach.

Biomarkers	**Control Enriched**	**T2DM Enriched**
Genus	*Eubacterium*	Alistipes
	Faecalibacterium	*Clostridium*
		Parabacteroides
		Subdoligranulum
Species	*Clostridiales sp. SS3/4*	*Escherichia coli*
	Eubacterium rectale	*Akkermansia muciniphila*
	Faecalibacterium prausnitzii	*Bacteroides intestinalis*
	Haemophilus parainfluenzae	*Bacteroides sp. 20_3*
	Roseburia intestinalis	*Clostridium bolteae*
	Roseburia inulinivorans	*Clostridium hathewayi*
		Clostridium ramosum
		Clostridium sp. HGF2
		Clostridium symbiosum
		Desulfovibrio sp. 3_1_syn3
		Eggerthella lenta

Chapter 5

Conclusion and Recommendations

Multiple studies have been conducted to determine the microbial signatures of Type II Diabetes Mellitus gut metagenomes. These studies typically focus on the taxonomic profiling of specific ethnicities utilizing amplicon sequencing because of its cost effectiveness and easy interpretation. A few meta-analyses have also been conducted employing multiple ethnicities but are based on either a read-based or an assembly-based approach alone. Such research limits the usability and reproducibility of the findings as population-specific biomarkers if applied to another population may lead to different results due to the genetic and technical variations. One such evidence is the discrepancies in the microbial biomarkers (composition as well as the abundance) and their association with the disease in different studies. Such discrepancies lead to false and non-reproducible results. Therefore, this study focused on utilizing an integrative approach employing both read-based and assembly-based methods to identify the differential biomarkers both at genus and species level. This study was unique as it employed samples from two different ethnicities i.e. American and Chinese with an aim to discover disease-specific biomarkers by overcoming the confounder effects, as well as the genetic and technical variations. It concluded that relying on anyone of the approach leads to a less accurate and limited result. Assembly-based approach determines

relatively lesser number of biomarkers possibly due to the limited size of the reference databases. It is a fact that despite advancements in the culture-independent methods, the size of the reference databases is still small. Thus, leaving most of the Metagenome Assembled Genomes unannotated. This limitation highlights the need for expansion in the size of the reference databases by advancing the current methodologies. Also, conducting a large-scale study employing more populations to identify disease-specific biomarkers utilizing both the approaches is required. Additionally, there is a need for methodological advancements especially for the identification of biomarkers by considering the integration of reads based and assembly-based approaches.

Bibliography

[1] J. Reed, S. Bain, and V. Kanamarlapudi, "A review of current trends with type 2 diabetes epidemiology, aetiology, pathogenesis, treatments and future perspectives," Diabetes, Metabolic Syndrome and Obesity, pp. 3567–3602, 2021.

[2] A. Sadagopan et al., "Understanding the role of the gut microbiome in diabetes and therapeutics targeting leaky gut: a systematic review," Cureus, vol. 15, no. 7, 2023.

[3] A. Berbudi, N. Rahmadika, A. I. Tjahjadi, and R. Ruslami, "Type 2 diabetes and its impact on the immune system," Curr Diabetes Rev, vol. 16, no. 5, pp. 442–449, 2020.

[4] S. Alam, M. K. Hasan, S. Neaz, N. Hussain, M. F. Hossain, and T. Rahman, "Diabetes Mellitus: insights from epidemiology, biochemistry, risk factors, diagnosis, complications and comprehensive management," Diabetology, vol. 2, no. 2, pp. 36–50, 2021.

[5] S. Bengmark, "Ecological control of the gastrointestinal tract. The role of probiotic flora," Gut, vol. 42, no. 1, pp. 2–7, 1998.

[6] F. Backhed, R. E. Ley, J. L. Sonnenburg, D. A. Peterson, and J. I. Gordon, "Host-bacterial mutualism in the human intestine," Science (1979), vol. 307, no. 5717, pp. 1915–1920, 2005.

[7] S. Altveş, H. K. Yildiz, and H. C. Vural, "Interaction of the microbiota with the human body in health and diseases," Biosci Microbiota Food Health, vol. 39, no. 2, pp. 23–32, 2020.

[8] B. Bakir-Gungor, O. Bulut, A. Jabeer, O. U. Nalbantoglu, and M. Yousef, "Discovering potential taxonomic biomarkers of type 2 diabetes from human gut microbiota via different feature selection methods," Front Microbiol, vol. 12, p. 628426, 2021.

[9] L. J. Brockley et al., "Sequence-based platforms for discovering biomarkers in liquid biopsy of non-small-cell lung cancer," Cancers (Basel), vol. 15, no. 8, p. 2275, 2023.

[10] J. Wirbel et al., "Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer," Nat Med, vol. 25, no. 4, pp. 679–689, 2019.

[11] H. Wen, "A comparison of pathogenesis of Diabetes in China and the United States," in E3S Web of Conferences, EDP Sciences, 2021, p. 2018.

[12] Y. Huang, S. Karuranga, B. Malanda, and D. R. R. Williams, "Call for data contribution to the IDF Diabetes Atlas 9th Edition 2019," Diabetes Res Clin Pract, vol. 140, pp. 351–352, 2018.

[13] U. Galicia-Garcia et al., "Pathophysiology of type 2 diabetes mellitus," Int J Mol Sci, vol. 21, no. 17, p. 6275, 2020.

[14] J. M. Rodríguez et al., "The composition of the gut microbiota throughout life, with an emphasis on early life," Microb Ecol Health Dis, vol. 26, no. 1, p. 26050, 2015.

[15] M. G. Dominguez-Bello et al., "Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns," Proceedings of the National Academy of Sciences, vol. 107, no. 26, pp. 11971–11975, 2010.

[16] A. L. Cunningham, J. W. Stephens, and D. A. Harris, "Gut microbiota influence in type 2 diabetes mellitus (T2DM)," Gut Pathog, vol. 13, pp. 1–13, 2021.

[17] W. Z. Gan, V. Ramachandran, C. S. Y. Lim, and R. Y. Koh, "Omics-based biomarkers in the diagnosis of diabetes," J Basic Clin Physiol Pharmacol, vol. 31, no. 2, p. 20190120, 2020.

[18] F. Padilla-Martinez, G. Wojciechowska, L. Szczerbinski, and A. Kretowski, "Circulating Nucleic Acid-Based Biomarkers of Type 2 Diabetes," International Journal of Molecular Sciences, vol. 23, no. 1. 2022. doi: 10.3390/ijms23010295.

[19] X. Chen and S. Devaraj, "Gut microbiome in obesity, metabolic syndrome, and diabetes," Curr Diab Rep, vol. 18, pp. 1–6, 2018.

[20] N. Segata et al., "Metagenomic biomarker discovery and explanation," Genome Biol, vol. 12, pp. 1–18, 2011.

[21] R. D. Leslie, J. Palmer, N. C. Schloot, and A. Lernmark, "Diabetes at the crossroads: relevance of disease classification to pathophysiology and treatment," Diabetologia, vol. 59, pp. 13–20, 2016.

[22] I. V Kononenko, O. M. Smirnova, A. Y. Mayorov, and M. V Shestakova, "Classification of diabetes. World Health Organization 2019. What's new?," Diabetes mellitus, vol. 23, no. 4, pp. 329–339, 2020.

[23] M. Yau, N. K. Maclaren, and M. A. Sperling, "Etiology and pathogenesis of diabetes mellitus in children and adolescents," Endotext [Internet], 2021.

[24] J. Reed, S. Bain, and V. Kanamarlapudi, "A review of current trends with type 2 diabetes epidemiology, aetiology, pathogenesis, treatments and future perspectives," Diabetes, Metabolic Syndrome and Obesity, pp. 3567–3602, 2021.

[25] A. V Hartstra, K. E. C. Bouter, F. Bäckhed, and M. Nieuwdorp, "Insights into the role of the microbiome in obesity and type 2 diabetes," Diabetes Care, vol. 38, no. 1, pp. 159–165, 2015.

[26] R. Goyal, M. Singhal, and I. Jialal, "Type 2 diabetes," StatPearls [Internet], 2023.

[27] L. Crudele, R. M. Gadaleta, M. Cariello, and A. Moschetta, "Gut microbiota in the pathogenesis and therapeutic approaches of diabetes," EBioMedicine, vol. 97, 2023.

[28] M. Gurung et al., "Role of gut microbiota in type 2 diabetes pathophysiology," EBioMedicine, vol. 51, 2020.

[29] N. N. Nam, H. D. K. Do, K. T. Loan Trinh, and N. Y. Lee, "Metagenomics: An effective approach for exploring microbial diversity and functions," Foods, vol. 12, no. 11, p. 2140, 2023.

[30] F. P. Breitwieser, J. Lu, and S. L. Salzberg, "A review of methods and databases for metagenomic classification and assembly," Brief Bioinform, vol. 20, no. 4, pp. 1125–1136, 2019.

[31] B. Kumar, J. Kinyua, and J. Kimotho, "Determination of microbial metagenomic markers of Type 2 Diabetes Mellitus (T2DM) in patients visiting South C Health Centre in Nairobi Kenya," Journal of Agriculture, Science and Technology, vol. 21, no. 2, pp. 83–95, 2022.

[32] M. Alshawaqfeh, A. Bashaireh, E. Serpedin, and J. Suchodolski, "Consistent metagenomic biomarker detection via robust PCA," Biol Direct, vol. 12, pp. 1–16, 2017.

[33] M. Alshawaqfeh, A. Bashaireh, E. Serpedin, and J. Suchodolski, "Reliable Biomarker discovery from Metagenomic data via RegLRSD algorithm," BMC Bioinformatics, vol. 18, pp. 1–17, 2017.

[34] I.-M. A. Chen et al., "IMG/M v. 5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes," Nucleic Acids Res, vol. 47, no. D1, pp. D666–D677, 2019.

[35] 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 Comput Biol, vol. 12, no. 7, p. e1004977, 2016.

[36] G. Ditzler, J. C. Morrison, Y. Lan, and G. L. Rosen, "Fizzy: feature subset selection for metagenomics," BMC Bioinformatics, vol. 16, no. 1, p. 358, 2015, doi: 10.1186/s12859-015-0793-8.

[37] M. B. Kursa and W. R. Rudnicki, "Feature Selection with the Boruta Package," J Stat Softw, vol. 36, no. 11 SE-Articles, pp. 1–13, Sep. 2010, doi: 10.18637/jss.v036.i11.

[38] M. Yassour et al., "Sub-clinical detection of gut microbial biomarkers of obesity and type 2 diabetes," Genome Med, vol. 8, no. 1, p. 17, 2016, doi: 10.1186/s13073-016-0271-6.

[39] S. Guo, H. Zhang, Y. Chu, Q. Jiang, and Y. Ma, "A neural network-based framework to understand the type 2 diabetes-related alteration of the human gut microbiome," iMeta, vol. 1, May 2022, doi: 10.1002/imt2.20.

[40] F. Padilla-Martinez, G. Wojciechowska, L. Szczerbinski, and A. Kretowski, "Circulating nucleic acid-based biomarkers of type 2 diabetes," Int J Mol Sci, vol. 23, no. 1, p. 295, 2021.

[41] A. Das and M. C. Schatz, "Sketching and sampling approaches for fast and accurate long read classification," BMC Bioinformatics, vol. 23, no. 1, p. 452, 2022, doi: 10.1186/s12859-022-05014-0.

[42] R. Huang, Y. Wang, D. Liu, S. Wang, H. Lv, and Z. Yan, "Long-Read Metagenomics of Marine Microbes Reveals Diversely Expressed Secondary Metabolites.," Microbiol Spectr, vol. 11, no. 4, p. e0150123, Aug. 2023, doi: 10.1128/spectrum.01501-23.

[43] J. Buttler and D. M. Drown, "Accuracy and Completeness of Long Read Metagenomic Assemblies.," Microorganisms, vol. 11, no. 1, Dec. 2022, doi: 10.3390/microorganisms11010096.

[44] L. Zhang et al., "Comparison Analysis of Different DNA Extraction Methods on Suitability for Long-Read Metagenomic Nanopore Sequencing.," Front Cell Infect Microbiol, vol. 12, p. 919903, 2022, doi: 10.3389/fcimb.2022.919903.

[45] V. Heidrich and L. Beule, "Are short-read amplicons suitable for the prediction of microbiome functional potential? A critical perspective.," iMeta, vol. 1, no. 3, p. e38, Sep. 2022, doi: 10.1002/imt2.38.

[46] S. Hampe, B. Batut, and P. Zierep, "Taxonomic Profiling and Visualization of Metagenomic Data," 2024.

[47] V. Sharma, N. Kumar, T. Prakash, and T. Taylor, "Fast and Accurate Taxonomic Assignments of Metagenomic Sequences Using MetaBin," PLoS One, vol. 7, p. e34030, Apr. 2012, doi: 10.1371/journal.pone.0034030.

[48] D. M. Portik, C. T. Brown, and N. T. Pierce-Ward, "Evaluation of taxonomic classification and profiling methods for long-read shotgun metagenomic sequencing datasets," BMC Bioinformatics, vol. 23, no. 1, p. 541, 2022, doi: 10.1186/s12859-022-05103-0.

[49] Q. Tran and V. Phan, "Assembling Reads Improves Taxonomic Classification of Species.," Genes (Basel), vol. 11, no. 8, Aug. 2020, doi: 10.3390/genes11080946.

[50] W. S. Pearman, N. E. Freed, and O. K. Silander, "Testing the advantages and disadvantages of short- and long- read eukaryotic metagenomics using simulated reads," BMC Bioinformatics, vol. 21, no. 1, p. 220, 2020, doi: 10.1186/s12859-020-3528-4.

[51] C. Milani et al., "METAnnotatorX2: a Comprehensive Tool for Deep and Shallow Metagenomic Data Set Analyses," mSystems, vol. 6, Jun. 2021, doi: 10.1128/mSystems.00583-21.

[52] G. Xu et al., "Combined assembly of long and short sequencing reads improve the efficiency of exploring the soil metagenome," BMC Genomics, vol. 23, no. 1, p. 37, 2022, doi: 10.1186/s12864-021-08260-3.

[53] J. Liu, S. Liu, Z. Yu, X. Qiu, R. Jiang, and W. Li, "Uncovering the gene regulatory network of type 2 diabetes through multi-omic data integration.," J Transl Med, vol. 20, no. 1, p. 604, Dec. 2022, doi: 10.1186/s12967-022-03826-5.

[54] D. S. H. Bell, "Combine and conquer: advantages and disadvantages of fixed-dose combination therapy," Diabetes Obes Metab, vol. 15, no. 4, pp. 291–300, 2013.

[55] E. Cersosimo, E. L. Johnson, C. Chovanes, and N. Skolnik, "Initiating therapy in patients newly diagnosed with type 2 diabetes: Combination therapy vs a stepwise approach.," Diabetes Obes Metab, vol. 20, no. 3, pp. 497–507, Mar. 2018, doi: 10.1111/dom.13108.

[56] N. Fierer et al., "Cross-biome metagenomic analyses of soil microbial communities and their functional attributes," Proceedings of the National Academy of Sciences, vol. 109, no. 52, pp. 21390–21395, 2012.

[57] D. Elmansy and M. Koyutürk, "Cross-population analysis for functional characterization of type II diabetes variants.," BMC Bioinformatics, vol. 20, no. Suppl 12, p. 320, Jun. 2019, doi: 10.1186/s12859-019-2835-0.

[58] L. Zhang et al., "Advances in Metagenomics and Its Application in Environmental Microorganisms," Front Microbiol, vol. 12, 2021, doi: 10.3389/fmicb.2021.766364.

[59] A. M. Shuikan, R. M. Alshuwaykan, and I. A. Arif, "The role of metagenomic approaches in the analysis of microbial community in extreme environment," in Life in Extreme Environments-Diversity, Adaptability and Valuable Resources of Bioactive Molecules, IntechOpen, 2023.

[60] J. L. Anders, M. A. M. Moustafa, W. M. A. Mohamed, T. Hayakawa, R. Nakao, and I. Koizumi, "Comparing the gut microbiome along the gastrointestinal tract of three sympatric species of wild rodents," Sci Rep, vol. 11, no. 1, p. 19929, 2021, doi: 10.1038/s41598-021-99379-6.

[61] M. K. Son, Y. Song, J. Chung, and H. S. Na, "Comparative Analysis of Healthy Gut Microbiota in German and Korean Populations: Insights from Large-Scale Cohort Studies," Microbiol Res (Pavia), vol. 15, no. 1, pp. 109–119, 2023.

[62] B. K. Meeks, K. A. Maki, N. J. Ames, and J. J. Barb, "Comparing Published Gut Microbiome Taxonomic Data Across Multinational Studies," Nurs Res, vol. 71, no. 1, pp. 43–53, 2022.

[63] M. U. R. Kayani, W. Huang, R. Feng, and L. Chen, "Genome-resolved metagenomics using environmental and clinical samples," Brief Bioinform, vol. 22, no. 5, pp. 1–20, 2021, doi: 10.1093/bib/bbab030.

[64] S. Chen, Y. Zhou, Y. Chen, and J. Gu, "Fastp: An ultra-fast all-in-one FASTQ preprocessor," Bioinformatics, vol. 34, no. 17, pp. i884–i890, 2018, doi: 10.1093/bioinformatics/bty560.

[65] P. Katara, Recent trends in "computational omics: concepts and methodology."

[66] J. A. O'Rawe, S. Ferson, and G. J. Lyon, "Accounting for uncertainty in DNA sequencing data," Trends in Genetics, vol. 31, no. 2, pp. 61–66, Feb. 2015, doi: 10.1016/J.TIG.2014.12.002.

[67] C. Quince, A. W. Walker, J. T. Simpson, N. J. Loman, and N. Segata, "Shotgun metagenomics, from sampling to analysis," Nat. Biotechnol., vol. 35, no. 9, pp. 833–844, Sep. 2017, doi: 10.1038/nbt.3935.

[68] E. Hauptfeld et al., "Integrating taxonomic signals from MAGs and contigs improves read annotation and taxonomic profiling of metagenomes", doi: 10.1038/s41467-024-47155-1.

[69] N. Segata, L. Waldron, A. Ballarini, V. Narasimhan, O. Jousson, and C. Huttenhower, "Metagenomic microbial community profiling using unique clade-specific marker genes," Nat. Methods, vol. 9, no. 8, pp. 811–814, Aug. 2012, doi: 10.1038/nmeth.2066.

[70] A. Blanco-Míguez et al., "Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4," Nature Biotechnology 2023 41:11, vol. 41, no. 11, pp. 1633–1644, Feb. 2023, doi: 10.1038/s41587-023-01688-w.

[71] H. Zafar and M. H. Saier, "Gut Bacteroides species in health and disease," 2021, doi: 10.1080/19490976.2020.1848158.

[72] E. Da, S. Morais, G. M. Grimaud, A. Warda, C. Stanton, and P. Ross, "Genome plasticity shapes the ecology and evolution of Phocaeicola dorei and Phocaeicola vulgatus," Scientific Reports |, vol. 14, p. 10109, 123AD, doi: 10.1038/s41598-024-59148-7.

[73] J. S. Glover, B. D. Browning, T. D. Ticer, A. C. Engevik, and M. A. Engevik, "Acinetobacter calcoaceticus is Well Adapted to Withstand Intestinal Stressors and Modulate the Gut Epithelium," Front Physiol, vol. 13, May 2022, doi: 10.3389/fphys.2022.880024.

[74] Y. Hong and J. D. Gu, "Bacterial anaerobic respiration and electron transfer relevant to the biotransformation of pollutants," International Biodeterioration and Biodegradation, vol. 63, no. 8. Elsevier Ltd, pp. 973–980, 2009. doi: 10.1016/j.ibiod.2009.08.001.

[75] S. Esquivel-Elizondo et al., "The Isolate Caproiciproducens sp. 7D4C2 Produces n-Caproate at Mildly Acidic Conditions From Hexoses: Genome and rBOX Comparison With Related Strains and Chain-Elongating Bacteria," Front Microbiol, vol. 11, p. 594524, Jan. 2021, doi: 10.3389/FMICB.2020.594524/BIBTEX.

[76] Y. K. Yeoh et al., "Prevotella species in the human gut is primarily comprised of Prevotella copri, Prevotella stercorea and related lineages," 123AD, doi: 10.1038/s41598-022-12721-4.

[77] M. Parsaei, N. Sarafraz, S. Y. Moaddab, and H. E. Leylabadlo, "The importance of Faecalibacterium prausnitzii in human health and diseases," 2021, doi: 10.1016/j.nmni.2021.100928.

[78] R. Higuchi et al., "Streptococcus australis and Ralstonia pickettii as Major Microbiota in Mesotheliomas," J. Pers. Med, 2021, doi: 10.3390/jpm11040297.

[79] M. Vacca, G. Celano, F. M. Calabrese, P. Portincasa, M. Gobbetti, and M. De Angelis, "The controversial role of human gut lachnospiraceae," Microorganisms, vol. 8, no. 4. MDPI AG, Apr. 01, 2020. doi: 10.3390/microorganisms8040573.

[80] E. Pasquereau-Kotula, M. Martins, L. Aymeric, and S. Dramsi, "Significance of Streptococcus gallolyticus subsp. gallolyticus association with colorectal cancer," Frontiers in Microbiology, vol. 9, no. APR. Frontiers Media S.A., Apr. 03, 2018. doi: 10.3389/fmicb.2018.00614.

[81] M. Gurung et al., "Role of gut microbiota in type 2 diabetes pathophysiology," EBioMedicine, vol. 51. Elsevier B.V., Jan. 01, 2020. doi: 10.1016/j.ebiom.2019.11.051.

[82] J. Wang et al., "A metagenome-wide association study of gut microbiota in type 2 diabetes," Nature 2012 490:7418, vol. 490, no. 7418, pp. 55–60, Sep. 2012, doi: 10.1038/nature11450.

[83] M. Gurung et al., "Role of gut microbiota in type 2 diabetes pathophysiology," EBioMedicine, vol. 51. Elsevier B.V., Jan. 01, 2020. doi: 10.1016/j.ebiom.2019.11.051.

[84] Z. Zhou, B. Sun, D. Yu, and C. Zhu, "Gut Microbiota: An Important Player in Type 2 Diabetes Mellitus," Frontiers in Cellular and Infection Microbiology, vol. 12. Frontiers Media S.A., Feb. 15, 2022. doi: 10.3389/fcimb.2022.834485.

[85] B. Bakir-Gungor, O. Bulut, A. Jabeer, O. U. Nalbantoglu, and M. Yousef, "Discovering Potential Taxonomic Biomarkers of Type 2 Diabetes From Human Gut Microbiota via Different Feature Selection Methods," Front Microbiol, vol. 12, no. August, pp. 1–16, 2021, doi: 10.3389/fmicb.2021.628426.

[86] L. Crudele, R. M. Gadaleta, M. Cariello, and A. Moschetta, "Gut microbiota in the pathogenesis and therapeutic approaches of diabetes," EBioMedicine, vol. 97, p. 104821, 2023, doi: 10.1016/j.ebiom.2023.104821.

[87] R. D. Hills, B. A. Pontefract, H. R. Mishcon, C. A. Black, S. C. Sutton, and C. R. Theberge, "Gut Microbiome: Profound Implications for Diet and

Disease," Nutrients 2019, Vol. 11, Page 1613, vol. 11, no. 7, p. 1613, Jul. 2019, doi: 10.3390/NU11071613.

[88] Y. Que et al., "Gut Bacterial Characteristics of Patients With Type 2 Diabetes Mellitus and the Application Potential," Front Immunol, vol. 12, Aug. 2021, doi: 10.3389/FIMMU.2021.722206/FULL.

[89] Y. Que et al., "Gut Bacterial Characteristics of Patients With Type 2 Diabetes Mellitus and the Application Potential," Front Immunol, vol. 12, Aug. 2021, doi: 10.3389/FIMMU.2021.722206/FULL.