CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

This chapter describes the research methodology used to achieve the objectives of this study. The description of study criteria i.e., research design, study location and research instruments are outlined. The information on the study participants, their selection criteria and the sampling method are also included. The instruments and procedures used for data collection are described and the methods used to analyze the data are discussed. Lastly, the methodology used for the systematic review is also explained.

3.2 Research Approach and Design

This is a case-control study whereby the study participants aged between 25 to 70 years old and from the three major ethnic groups (Malay, Chinese and Indian) were included. The cases were defined as participants who attended the diabetes clinic, Klinik Kesihatan Ampang and were diagnosed with T2DM for at least 1 year. The controls were confirmed non-diabetic subjects attending the outpatient department (OPD) for a non-communicable disease screening (NCD) in Klinik Kesihatan Ampang and had FPG of <6.1 mmol/L.

3.3 Study Location/Setting

The study was conducted in Klinik Kesihatan Ampang, Selangor as this was one of the main clinics in Ampang to be attended by individuals from various ethnic groups. The research laboratory in the Faculty of Medicine and Health Science at Universiti Sains Islam Malaysia (FPSK USIM, Nilai) was used to analyze the samples collected.

3.4 Target Population

The diabetic and non-diabetic adults in the multi-ethnic population around Ampang in Selangor, Malaysia.

3.5 Study Population

The study population consisted of patients attending the diabetes clinic and outpatient department (OPD) in Klinik Kesihatan Ampang, Selangor. Hence, they were either confirmed diabetic or non-diabetic and belonged to one of the three major ethnic groups in Malaysia.

3.6 Sampling Method and Sample Size

Sampling method known as probability sampling with systemic stratified sampling will be used. The sample size was calculated using online Open Epi calculator, based on data from Kannan et al. (2017), on the relative abundance of bacterial phyla *Firmicutes* (diabetics 21%, controls 51%) and *Bacteroidetes* (diabetics

73%, controls 46%). With the confidence level set at 95% and power of study set at 80%, the calculated sample size using Fleiss is 37. Additional 20% non-response rate was added bringing a total of 45 per arm and 90 subjects overall. Faecal samples will be collected from 15 T2DM and 15 nonDM Malaysians of each Malay, Chinese and Indian ethnicity making the total number of samples, 90 i.e. [(15+15) *3 = 90].

3.7 Inclusion and Exclusion Criteria

3.7.1 Inclusion and Exclusion Criteria for T2DM Participants

Inclusion criteria:

i. Diagnosed with T2DM for less than 10 years.

ii. FPG level of > 7.0 mmol/L

iii. Following a medication regime for T2DM during the time of study.

Exclusion criteria:

i. Intake of antibiotics or probiotics for the past 3 months

ii. Acute infections [fever, diarrhea, urinary tract infection (UTI), respiratory infections]

iii. Chronic illness or complications (dyslipidemia, chronic kidney disease, cataract, heart-related disease, history of stroke, gastrointestinal disease, cancer, and recent surgery)

iv. Involved in dietary interventions.

v. Wheelchair bound (may have difficulty in collecting stool without contamination)

vi. Pregnant or lactating

3.7.2 Inclusion and Exclusion Criteria for NonDM Participants

Inclusion criteria:

i. Absence of T2DM diagnosis

ii. FPG of < 6.1 mmol/L

iii. Individuals who came for NCD (non-communicable disease) screening.

Exclusion criteria:

i. Intake of antibiotics or probiotics for the past 3 months

ii. Acute infections [fever, diarrhea, urinary tract infection (UTI), respiratory infections]

iii. Chronic illness or complications (dyslipidemia, chronic kidney disease, cataract, heart-related disease, history of stroke, gastrointestinal disease, cancer, and recent surgery)

iv. Involved in dietary interventions.

v. Wheelchair bound (may have difficulty in collecting stool without contamination)

vi. Pregnant or lactating

3.8 Methodology Flowchart

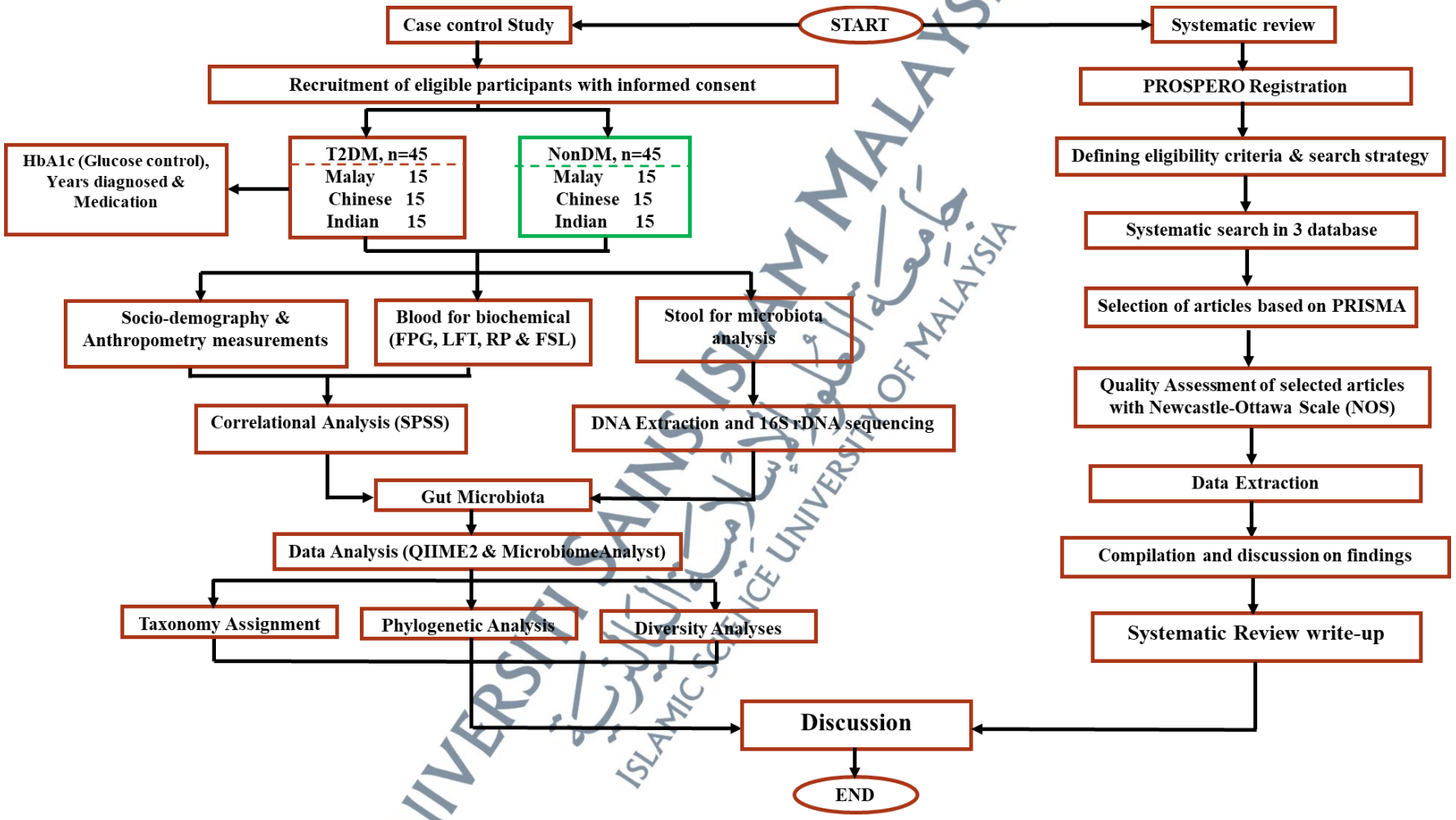

Figure 3.1 Methodology Flowchart.

3.9 Research Instruments

3.9.1 Questionnaire

A questionnaire was used in this study to collect information on the subject's socio-demography (age, date of birth, gender and ethnicity). This part also consisted of the subject's history of smoking, if the first-degree family had a history of diabetes, and the subject's other health problems if any.

3.9.2 Anthropometry

The weight and height of the participants were recorded in a mechanical weighing scale (ZT-120).

3.9.3 Blood Request Form

A completed request form for blood tests was provided by the doctor to all the study subjects (refer to Appendix B). The blood samples were collected in two ethylenediamine tetra-acetic acid (EDTA) tubes (2ml/tube) with the help of the medical personnel in the clinic. The blood tubes were immediately sent to the laboratory unit at Klinik Kesihatan Ampang and processed with hematology analyzer (Beckman Coulter).

3.9.4 Stool Collection Kit

The stool kit was provided for all the study subjects. The stool collection kit contained sterile gloves, a faeces catcher (a strip of fecal collection paper that fits on a

toilet seat) (Zymo Research) and a 200ml stool container with a screw-tight lid and spoon.

3.9.5 Lab-Work Tools and Machines

The stool samples were homogenized by using Mini Beadbeater-1 (Biospec Products). The DNA extraction was carried out with the ZymoBiomics DNA Miniprep extraction kit (Zymo Research). The concentration of DNA was quantified using nanophotometer (Implen Nanophotometer) and fluorometric quantification (iQuant Broad Range dsDNA Quantification Kit).

3.9.6 Data Analysis Software

A Linux based software named QIIME2 (Qualitative Insights into Microbial Ecology) (version 2021.8) was used to conduct downstream analyses with the data obtained from sequencing. The metadata file used in QIIME2 was validated with KEEMEI, a Google sheet add-on that assessed the validity of bioinformatics file formats (Rideout et al., 2016). The taxonomy diagrams and visualizations were constructed by using QIIME2 and Microsoft Excel. The R software (version 4.0.2) in MicrobiomeAnalyst was used to construct rarefaction curve and for diversity analyses (Dhariwal et al., 2017). The SPSS (version 25) was used for correlation and statistical analyses.

In systematic review, the electronic databases i.e. PMC-NCBI (PubMed Central by National Centre for Biotechnology Information), MEDLINE Complete, and CINAHL (EBSCO Host) were used for the systematic search of articles.

3.10 Operational Definition

Variables	Operational Definition
T2DM	Individuals with FPG ≥ 7.0 mmol/L (Clinical Practice Guidelines, 2015)
NonDM	Individuals with FPG < 6.1 mmol/L (Clinical Practice Guidelines, 2015).
Ethnicity	The ethnic class of individuals as reported by the national registration identification card (NRIC).
Gut Microbiota Composition	The bacterial phylum and genus found in all samples based on ASV table obtained from 16S rDNA amplicon sequencing.
Alpha-diversity	The species richness (presence or absence of species by Observed ASVs and Chao1 indices) and evenness (frequency of species occurrence by Pielou's evenness index) as well as both richness and evenness (by Shannon index) within samples (Chong et al., 2020)
Beta-diversity	The differences in gut microbiota structure between the study groups based on unweighted UniFrac (absence or presence of ASVs), weighted UniFrac (differences in relative abundance) as well Bray-Curtis dissimilarity matrix (analyses dissimilar microbial species) (Chong et al., 2020).
Abundance	The percentage of relative abundance of specific bacterial phylum & genus from the total number of bacteria found across all samples (Young et al., 2008)
Clinical Characteristics	The anthropometry measurements (BMI, height and weight), demographic (age), diabetic profile (FPG) as well as the biochemical parameters (LFT, RP and FSL) of the study participants.
PreDM	Individuals with either isolated impaired fasting glucose (IFG) (≥6.1 to 6.9 mmol/L), isolated impaired glucose tolerance (IGT) (with FPG : <7.0 mmol/L and OGTT: ≥7.8 and <11.1mmol/L) (WHO, 2006).
NewDM	Individuals with OGTT [FPG (> 7.0 mmol/L), 2-HPP (≥11.1 mmol/L)] or with HbA1c from ≥ 6.5% and have not begun anti-diabetic medications (Clinical Practice Guidelines, 2015).

3.11 Definition and Normal Range of Variables Measured in this Study

Variables	Definition	Normal Range
BMI	A health index that uses an adult's weight and height to estimate the amount of body fat in an individual.	According to Asian-Pacific BMI (WHO Expert Consultation, 2004) Underweight <18.5 kg/m^2 Normal weight:18.5–23 kg/m^2 Overweight: 23 – 27.5 kg/m^2 Obese ≥ 27.5 kg/m^2
FPG	The blood glucose level measured after a fasting period of 8 hours.	< 6.1 mmol/L
TP	A measurement of total protein found in the blood.	66-87 g/L
Alb	A measurement of the protein albumin in blood	35-50g/L
ALP	These enzymes which are mostly found in liver is an indicator of the liver health	35-105 U/L
ALT		<31 U/L
TB	This is a measurement of both direct and indirect bilirubin found in blood.	5-17 µmol/L
Ur	The amount of urea measured in blood reflects the kidney health	1.7 – 8.3 mmol/L
Na	These are electrolytes measured in the blood to assess kidney function	135-145 mmol/L
K		3.5 – 6.0 mmol/L
Cl		95 – 110 mmol/L
Cr	A by-product of energy-producing muscles found in the blood reflects the efficiency of filtration system in kidney.	44 – 80 µmol/L
TC	A measure of both HDL and LDL levels in blood	< 5.7 mmol/L
TG	A measure of the amount of fat in the form of triglycerides found in blood.	< 1.7 mmol/L
HDL-C	Known as 'good cholesterol' measured in blood.	> 1.4 mmol/L
LDL-C	Known as 'bad cholesterol' measured in blood.	< 3/9 mmol/L
HDL/TC	The ratio of TC and HDL levels reflects lipid metabolism in body.	> 25%
HbA1c	A glycated haemoglobin measured in blood to estimate blood sugar level and glucose control for the past 3 months	≤ 6.5%
Glucose control (based on HbA1c, %)		Good glucose control ≤ 6.5% Poor glucose control ≥ 6.5%

Source: (Clinical Practice Guidelines, 2015)

3.12 Data and Samples Collection Method

Participants were enrolled in this study between November 2019 to October 2020. Participants were briefed on the study purpose and design and provided a signed consent form (refer Appendix C). The anthropometric and socio-demographic details of the participants were measured and recorded in a questionnaire (refer Appendix D). Additionally for the diabetic patients, period since diagnosis of T2DM and the medication taken were also recorded. The participants were given an appointment date for blood taking after overnight fasting. Blood samples were taken and clinical parameters i.e., liver function test (LFT), renal profile (RP), fasting serum lipid (FSL), and FPG were measured in the Klinik Kesihatan Ampang laboratory. The HbA1c test was also included only for T2DM participants. The participants were also given a stool collection kit and explained the correct procedure for proper stool collection to minimise contamination from the environment. They were advised to return the tubes within two hours of collection. The collected samples were transported in an icebox to the laboratory and stored at -80ºC before further processing.

3.13 Genomic DNA Extraction

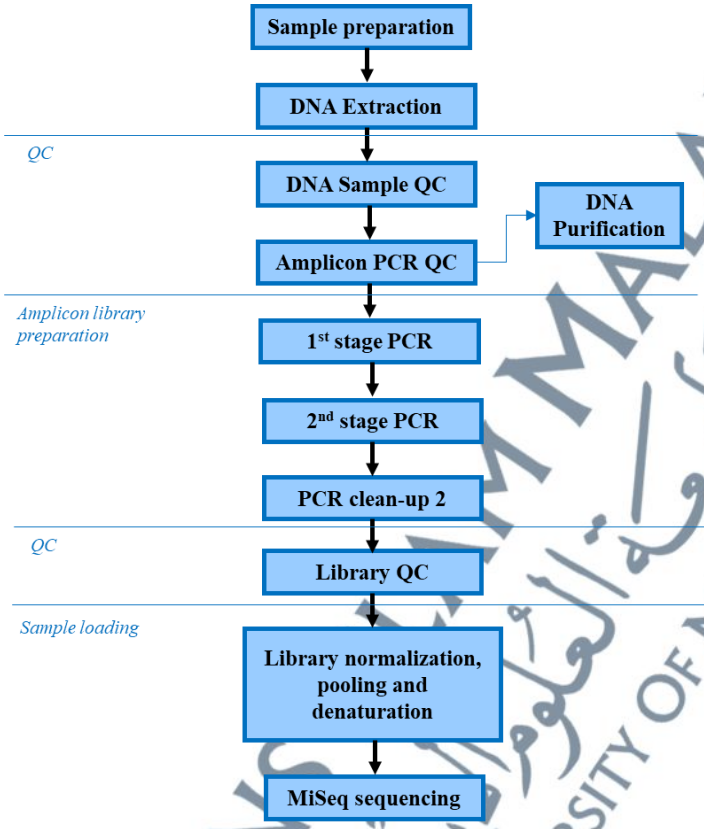

Figure 3.2 Flowchart for Gut Microbiome Analysis.

The stool samples were processed further for gut microbiome analysis, as shown in Figure 3.2. Total faecal genomic DNA (gDNA) was extracted from 90 faecal samples. A couple of extractions were also performed using only water instead of faecal material, to serve as negative controls. The thawed faecal samples (0.20g) were transferred into ZR BashingBead lysis tubes with 750µL of Lysis Solution. Then, these tubes were homogenized by bead beating for 1 minute at 4800rpm [For optimization of bead beating time and speed, DNA from two stool samples were extracted with varying bead beating times (0 – 3 min) and speeds (0 – 4800 rpm). The best bead beating time

and speed were determined based on optimum concentration (≥ 20 ng/ul) and minimal smearing in agarose gel electrophoresis, of the DNA extracted.]

The homogenized faecal suspension was centrifuged at 10,000 x g for 1 min and the supernatant (400 µL) was applied to the Zymo-Spin III-F Filter and centrifuged at 8,000 x g for 1 min. The ZymoBIOMICS DNA Binding Buffer (1.2 mL) was added to the filtrate in the Collection Tube. Later, 800 µL of the mixture was applied to the Zymo-Spin IICR Column and centrifuged at 10,000 x g for 1 min. This step was repeated after discarding flow-through. DNA Wash Buffer 1 (400 µL) was applied to the Zymo-Spin IICR Column centrifuged at 10,000 x g for 1 minute. Later DNA Wash Buffer 2 (700 µL) was added to the Zymo-Spin IICR Column centrifuged at 10,000 x g for 1 minute and this step is repeated with DNA Wash Buffer 2 (200 µL). Next, 100 µL of DNase/RNase Free Water was applied to the column and incubated for 1 min, then centrifuged at 10,000 x g for 1 minute to elute DNA. Eluted DNA was applied to the Zymo-Spin III-HRC Filter, centrifuged at 8,000 x g for 3 min, and stored at -20 ˚C until processing. The gDNA extracted were expected to have a concentration of at least 20 ng/ul and DNA purity of A_{260}/A_{280} (1.7 – 1.9) and A_{260}/A_{230} (1.8 and above). For samples with low concentration or purity, the extraction was repeated by first diluting the stool sample (0.20g) in 1ml of PBS and centrifugation at 10,000 x g for 1 min to remove any impurities. The washed stool pellets were then used to proceed with DNA extraction according to the protocol.

3.14 DNA Quality Control and Sequencing

A total of 90 samples of gDNA and two negative controls were submitted to Apical Scientific Sdn.Bhd. for 16S rDNA sequencing. The following gDNA quality

control (QC) along with the 16S rDNA sequencing were performed by the service provider (Appendix E).

Firstly, the quality of gDNA was checked on 1% TAE agarose gel. Aliquots of 1 µl gDNA were run on 1% TAE agarose gel at 100V for 60 minutes. This was run along a 1kb ladder and a positive control (50 ng of the bacterial gDNA. Then, the gDNA bands from the gel were checked for signs of degradation or contamination for each sample. The concentration of gDNA was again measured using a nanophotometer and fluorometric quantification (refer Appendix F).

The purified gDNA was then proceeded to amplicon PCR QC and amplified using locus-specific sequence primers, 341F/806R. The gDNA that passed the amplicon PCR QC were eligible for library preparation. The amplicon library preparation was done with 2-Step PCR, according to Illumina's 16S metagenomic library preparation guideline (Illumina, 2013). The libraries that passed the library QC were normalized and pooled according to the protocol recommended by Illumina. This is followed by sequencing using the MiSeq platform with MiSeq Reagent Kit version 3. This generated more than 50 million reads in a pair-end format with a 300bp read length (Illumina, 2013).

3.15 Data Analysis

3.15.1 Sequence QC

The 300 bp paired-end (PE) raw reads obtained from service provider was demultiplexed, i.e., split into two individual fastq files (forward and reverse reads) for each sample. Then, the sequences were aligned, merged and denoised using Divisive Amplicon Denoising Algorithm 2 (DADA2) pipeline (version 1.14) (Callahan, 2020).

These steps are crucial to remove and/or correct read errors, remove low-quality regions, chimeric errors and merge denoised paired-ends reads (Nearing et al., 2018). The quality control of the DADA2 output were analysed and visualized using MultiQC (Ewels et al., 2016) (Figure 3.3). This is a web-based software that analyses the quality of sequencing data and highlights any potential problems in data prior to dowstream analyses. The end result is an amplicon sequence variant (ASV) table which stores the number of times each ASV was observed in each sample.

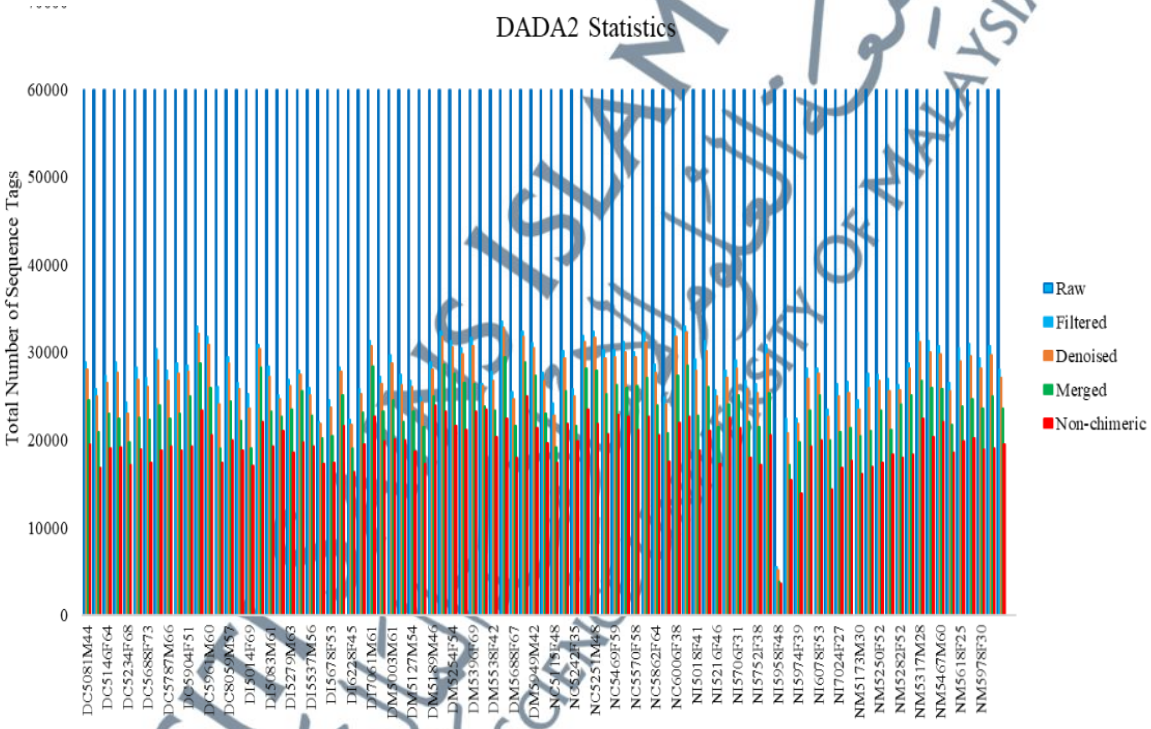

Figure 3.3 Sequence Quality Control using DADA2 Pipeline.

3.15.2 Taxonomy Assignment

The ASV table were used for further downstream analyses of taxonomic classification and calculation of the relative frequency of ASVs in each sample. SILVA, a 16S rRNA gene reference database was used to prepare a QIIME2 (Bolyen et al., 2019) compatible SILVA-based 16S-taxonomy classifier (Quast et al., 2013; Robeson et al., 2021). This trained classifier, based on the amplicon-specific reference sequence files, identified and clustered the similar ASV representative sequences (obtained from the study population) at 99% similarity for each taxonomy level. This created a taxonomy file with the assigned taxonomy labels for ASV sequences. Then, this taxonomy file, the ASV frequency table and a metadata file (data from study participants) were used in QIIME2 to visualize the taxonomic distribution in Phylum and Genus level in all study participants.

In respective study groups by diabetes status and/or ethnicity, the percentage abundance of phyla and the genera with more than 1% abundance were visualized by using Microsoft Excel. The *Firmicutes*/*Bacteroidetes* (F/B) ratio was calculated by using a standard formula: (A/B : B/B), where A is the relative frequency of *Bacteroidetes* and B is the relative frequency of *Firmicutes* (Gaike et al., 2020). The F/B ratio was calculated for each sample and the average ratio for each study group was tabulated.

3.15.3 Diversity Analysis

The MicrobiomeAnalyst, a web-based software that analyses and visualise the microbiome data was used for data normalization and diversity analyses (Dhariwal et al., 2017). This software generated a rarefaction curve that displayed each ASV's sequence length for all the samples. Based on the minimum ASV sequence that could retain most of the samples, the suitable sequence length (13, 900 reads/sample) was selected for data normalization. This step excluded one sample with the lowest sequence count of 3649 and only 89 samples were used further for diversity analyses.

The diversity analyses includes both alpha and beta diversity which followed the protocol available in QIIME 2 software (Bolyen et al., 2019). Alpha diversity analysed gut microbiota richness (presence or absence of species in a sample) by using Observed species and Chao1. The gut microbiota evenness (distribution of species abundance in a sample) is analysed using Pielou's eveness index while both species richness and abundance (the frequency of species occurrence in a sample) are analysed by using the Shannon index. Beta-diversity or the analysis of gut bacteria structure between the study groups were measured by using three dissimilarity matrices: weighted UniFrac, unweighted UniFrac and Bray-Curtis. These matrices were used to construct a Principal Coordinate Analysis (PCoA) plot. This plot highlighted the clustering patterns in bacterial communities between the study groups (Chong et al., 2020).

3.15.4 Statistical Analysis

The participant characteristics based on diabetes status and/or ethnicity were analyzed using the SPSS v25 (McCormick et al., 2020). The continuous data was expressed as mean ± standard deviation (SD) for normally distributed data and median (interquartile range, IQR) for non-normally distributed (skewed) data while categorical data was expressed in frequencies, n. The parametric test, ANOVA (Analysis of Variance) [post-hoc test (Bonferroni)] was used for normal continuous data while the non-parametric test, Kruskal-Wallis test (pair-wise comparison with Mann-Whitney U test) were used for non-normal continuous data. In both normal and non-normal data, simple logistic regression analysed the continuous variables while Chi-square test analysed the categorical variables (Ramakrishna, 2016).

Each alpha diversity metric was analyzed with Student's t test for normal data and Mann Whitney U test for non-normal data by using SPSS (Dhariwal et al., 2017). The beta diversity measures between the study groups were analysed by using PERMANOVA (per Mutational Multivariate Analysis of Variance) test by using QIIME2 (Hall et al., 2018). The taxonomic data was expressed in prevalence (%) and abundance (%). The taxonomic differences among study groups were analysed by using Kruskal-Wallis or Mann-Whitney U test. The p-values were corrected using False Discovery Rate (FDR) with the Benjamin-Hochberg method to account for multiple comparisons. Lastly, Spearman correlation was used to test the association between anthropometric, demographic and clinical characteristics in all study participants (Stehlik-Barry et al., 2017). Spearman correlation test was done by using SPSS and the heatmap was created in Microsoft Excel. A value of adjusted $p < 0.050$ (FDR) was considered statistically significant.

3.16 Research Ethics

i. Study was conducted according to good clinical practice guidelines (GCP)

ii. Informed consent and written consent was taken

All the study participants were briefed on the study details. The sample was collected upon receiving full consent from the study participants. The protection of privacy of the study participants providing stool samples was ensured. This was done by labelling each sample with a specific ID instead of the study participant's name. The anonymity of participants was maintained for the entire course of this study.

iii. Invasive protocol

The study participants were made aware of the blood taking procedures which is the only invasive procedure in this study.

iv. Participants were offered if interested in knowing their microbiota data

The participants were informed that the microbiome data would be enclosed to them, if they were interested.

v. Ethic approval

The ethical approval from this study was provided by the Ministry of Health, Malaysia [NMMR-18-3688-44808(IIR)]

vi. Pandemic considerations (SOP)

The standard operating procedures (SOP) were maintained, i.e., social distancing during participant selection, interview session and sample collection. Any study participants who exhibited the symptoms for COVID-19 were not included in the study.

3.17 Methodology for Systematic Review

A systematic review of observational studies was carried out according to a protocol published in the International Prospective Register of Systematic Reviews (PROSPERO) (CRD42020160458, 10/7/2020) reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Moher et al., 2009). This systematic review evaluated and summarized the microbiota diversity in prediabetes individuals (preDM) and newly diagnosed T2DM individuals (newDM) as compared to non-diabetic individuals (nonDM) (Appendix A).

A systematic search of published literature was performed using electronic databases [PubMed Central by National Centre for Biotechnology Information (PMC-NCBI), MEDLINE Complete, and CINAHL (EBSCO Host)] from inception up to February 2021. Search strategies were developed based on the keywords related to "type 2 diabetes", "prediabetes", "newly-diagnosed" and "gut microbiota". There were 18 chosen articles that were assessed with the Newcastle Ottawa scale and the data were extracted and analysed further. A standard form was used to extract the data from included studies. The primary outcomes were gut microbial abundance and differences between study groups at the phylum, class, order, family, and genus taxonomic ranks. Species were grouped into their respective genus. The secondary outcomes were clinical characteristics, dietary intake, or other parameters measured and their correlation with the gut microbial composition.

3.18 Conclusion

This chapter focused on the research methodology adapted in this study. The research design and the sampling of study population were discussed. This is followed by the data collection procedure and sample collection. The collected samples were analysed by using a series of laboratory procedures. The data from these procedures are subjected to data analysis. Lastly, the methodology adapted in the systematic review was discussed. The result and findings of this chapter are the outline for the next chapter.