BIOMARKER DISCOVERY UTILIZING BIG DATA: THE CASE OF DIABETES IN UNITED ARAB EMIRATES

by

Bayan Hassan Banimfreg

A Dissertation Presented to the Faculty of the
American University of Sharjah
College of Engineering
in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy in
Engineering Systems Management

Sharjah, United Arab Emirates

May 2022

Declaration of Authorship

I declare that this dissertation is my own work and, to the best of my knowledge and belief, it does not contain material published or written by a third party, except where permission has been obtained and/or appropriately cited through full and accurate referencing.

Signature: Bayan Hassan Banimfreg

Date: 5/6/2022

Approval Signatures

We, the undersigned, approve the PhD Dissertation written by: Bayan Hassan Banimfreg

Dissertation Title: Biomarker Discovery Utilizing Big Data: The Case of Diabetes in United Arab Emirates

Date of Defense: April 20, 2022

Name, Title and Affiliation	**Signature**
Dr . Abdulrahim Shamayleh Assistant Professor Department of Industrial Engineering Dissertation Advisor	
Dr . Hussam Alshraideh Associate Professor Department of Industrial Engineering Dissertation Co-advisor	
Dr. Ayman Alzaatreh Associate Professor Department of Mathematics and Statistics Dissertation Committee Member	
Dr . Mahmoud Awad Associate Professor Department of Industrial Engineering Dissertation Committee Member	
Dr . Nelson Soares Assistant Professor Department of Medicinal Chemistry Dissertation Committee Member	
Dr . Zied Bahroun Associate Professor Department of Industrial Engineering Dissertation Examiner	
Dr . Omar Ashour Associate Professor Department of Industrial Engineering Dissertation Examiner	

Accepted by:
Dr. Mohamed El-Tarhuni
Vice Provost for Graduate Studies
Office of Graduate Studies

Acknowledgment

First and foremost, I would like to thank Dr. Abdulrahim Shamayleh and Dr. Hussam Alshraideh for their supervision and continued support throughout this journey. I have learned so much from you about how to be a researcher, and it is a solid foundation that I will be able to build on for years to come. And finally, thank you for patiently helping me improve my scientific writing skills.

I would also like to thank Dr. Nelson Soares for his help and commitment toward achieving our work. It has also been a great pleasure collaborating with the Sharjah Institute of Medical Research (SIMR). Thanks to everyone involved in this research, whether minor or significant contribution.

My father and mother, I pray to God to grant you health, prosperity, and happiness. You are both my source of love, encouragement, and eternal happiness. My brothers and sisters, thanks for always being there for me, and without you, I never would have gotten this far.

Joory, Suhail, and Kenan, my little angels, I am unbelievably grateful for your ever-lasting and unconditional love. Thank you for being patient and resilient kids. I could not express my love and gratitude to my husband, Mohammad. Thanks for putting up with being married to a graduate student for so long.

Thank you to the American University of Sharjah for having me as a member of your highly valued environment. My gratitude for the Faculty Research Group (FRG) funding, Teaching, and Research Assistantships opportunities to undertake my Ph.D. studies. Thank you for supporting me during this journey.

Lastly, but not least, I would like to thank my fellow graduate student friends, who have made the years fly by with fun, and close friends, who have always supported my educational quests.

Abstract

Diabetes Mellitus (DM) received substantial attention for exploring its mechanism as expected to be the seventh primary reason for death worldwide by 2030. The hallmark of DM leads to damaging effects on many organ systems, mainly the cardiovascular, ophthalmic, and renal systems. The number of adults with DM to reach 95 million by 2030 and 136 million by 2045 in the Middle East and North Africa region. Type 2 diabetes (T2DM) is the most common type of DM, accounting for around 90% of diabetes cases. T2DM is a multifactorial chronic metabolic disease caused by genetic and non-genetic factors resulting from an imbalance between energy intake and output and other lifestyle-related factors. However, the detailed understanding of T2DM etiology is still limited. As the focus of this work is the metabolomic derived biomarker discovery, a non-targeted metabolomics experiment using liquid chromatography with tandem mass spectrometry (LC-MS/MS) is conducted to explore the metabolic profile of diabetic Emirati Citizens to uncover potential novel diabetes biomarkers through big data analytics. The study is twofold: in the first part, a comprehensive analysis is performed to reveal the profiling metabolites of diabetic Emirates compared to healthy ones. Blood samples of 50 diabetic Emiratis versus 42 healthy were utilized to investigate for differential metabolites. In the second part, a metabolomic study of patients with diabetic kidney disease against dialysis non-diabetics patients was conducted to uncover their distinct biomarkers. Blood samples of 11 dialysis diabetics and 25 dialysis non-diabetic were used to reveal potential biomarkers. A great panel of potential differential metabolites was identified among diabetic and non-diabetic Emirates. The identified metabolites were sorted into classes, including Tryptophan and Purines. Several potential biomarkers and their related pathways were pinpointed among dialysis patients, including Tyrosine metabolism-related metabolite and 3,4-Dihydroxymandelic acid. These studies provide detailed coverage of blood metabolic changes related to T2DM in the Emirati population. The results of this work are mainly consistent with similar international studies, with a few added biomarkers reflecting the region-specific health profile. The worldwide consensus on common metabolites encourages the clinical trials of novel biomarkers that could expedite the treatment process for diabetics. Monitoring and managing diseases might move medicine from a therapeutic model to a prevention model.

Keywords: Metabolomics; Biomarker discovery; Diabetes; Pathway analysis Liquid chromatography with tandem mass spectrometry; United Arab Emirates.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AAAs	Aromatic Amino Acids
ALS	Asymmetric Least Squares
AUC	Area Under the Curve
BCAA	Branched-Chain Amino Acids
BioDiscML	Biomarker Discovery by Machine Learning
BMI	Body Mass Index
BRCA1	Breast Cancer Type 1
BRENDA	The BRaunschweig ENzyme Database
cAMP	Cyclic Adenosine Monophosphate
CE–MS	Capillary Electrophoresis–Mass Spectrometry
CHD	Coronary Heart Disease
ChEBI	Chemical Entities of Biological Interest
CKD	Chronic Kidney Disease
COW	Correlation-Optimized Warping
CSF	Cerebrospinal fluid
DD	Dialysis Diabetics
DKD	Diabetic Kidney Disease
DM	Diabetes Mellitus
DNA	Deoxyribonucleic Acid
DND	Dialysis non-Diabetic Patients
ENCODE	Encyclopedia of DNA Elements
ESRD	End-Stage Renal Disease
FDR	False Discovery Rate
FIA	MS/MS-Flow injection analysis tandem mass spectrometry
FQM	Feature Quantification Matrix
FT	Fourier transform
GC-MS	Gas Chromatography Coupled to Mass Spectrometry
GLUT4	Glucose Transport of Protein 4
GMD	Golm Metabolome Database
HbA1C	Glycated Hemoglobin
HCA	Hierarchical Clustering Analysis
HER	Electronic Health Records

HMDB	Human Metabolome Database
IDF	International Diabetes Foundation
IFG	Impaired Fasting Glucose
IGT	Impaired Glucose Tolerance
InsR	Insulin Receptor
IR	Insulin resistance
IRS	Insulin Receptor Substrate
KEGG	Kyoto Encyclopedia of Genes and Genomes
KNN	k-nearest Neighbors
LC-MS	Liquid Chromatography with Mass Spectrometry
LC-MSMS	Liquid Chromatography with Tandem Mass Spectrometry
LeapR	The Layered Enrichment Analysis of Pathways
m/z	Mass-to-Charge Ratio
MassTRIX	Mass TRanslator into Pathways
MBROLE	Metabolites Biological Role
MENA	The Middle East and North Africa
MENDA	Metabolite Network of Depression Database
MPEA	Metabolite Pathway Enrichment Analysis
MS	Mass Spectrometry
NCDs	Noncommunicable Diseases
ND	Non-diabetics
NMR	Nuclear magnetic resonance spectroscopy
ORA	Over-Representation Analysis
OSC	Orthogonal Signal Correction
PA	Pathway Analysis
PANEV	PAthway NEtwork Visualizer
PAPi	Pathway Activity Profiling
PCA	Principal Component Analysis
PCs	PhosphatidylCholines
PI3K	Phosphatidylinositol 3-Kinase
PTB	Pathway-Topology Based
RaMP	The relational database of Metabolomics Pathways
RBE	Robust Baseline Estimation
RNA	Ribonucleic Acid

RNS	Reactive Nitrogen Species
ROS	Reactive Oxygen Species
SAM	Significance Analysis of Microarrays
SCFAs	Short-Chain Fatty Acids
SDBS	Spectral Database System
SIMR	Sharjah Institute of Medical Research
SMPDB	The Small Molecule Pathway Database
SOMs	Self-Organizing Maps
STOCSY	Statistical Total Correlation Spectroscopy
SVMs	Support Vector machines
T1DM	Type 1 Diabetes
T2DM	Type 2 Diabetes
TCGA	The Cancer Genome Atlas
UAE	United Arab Emirates
UPLC	MS-Ultraperformance Liquid Chromatography Coupled to Mass Spectrometry
VAT	Visceral Adipose Tissue
VMH	Virtual Metabolic Human
WHO	World Health Organization
XML	Extensible Markup Language

Chapter 1. Introduction

1.1 Diabetes Definition

Chronic diseases, defined as noncommunicable diseases (NCDs), are complex disorders that tend to be of long duration and result from a combination of genetic, physiological, environmental, and behavioral factors. Chronic diseases such as Diabetes Mellitus (DM) are currently considered the leading cause of morbidity and mortality globally, along with an alarming growth in developed and developing nations. DM is a chronic disease of the metabolic system, defined by chronic high blood sugar level hyperglycemia, which might severely damage the entire body. Continuing high blood sugar can harm the eyes, blood vessels, kidneys and causes skin infections and slow healing of cuts and sores. The disease diagnosis might already present chronic or long-term DM difficulties in individuals with T2DM. DM happens when the body cannot generate enough insulin or cannot use insulin successfully. Insulin is a hormone produced in the pancreas that permits glucose from food to penetrate the body's cells. The latter translates glucose into energy required by muscles and tissues to work. A person with DM does not absorb glucose correctly, and glucose remains circulating in the blood, a condition recognized as hyperglycemia, hurting body tissues over time. Overall, it will cause long-term consequences that substantially worsen the quality of life. DM has two major classes: type 1 (T1DM) and type 2 (T2DM). However, about 90-95% of people with diabetes have T2DM [1].

T1DM primarily occurs due to having a genetic predisposition. Therefore, it mostly happens in children, adolescents, and adults. In comparison, the reasons for or direct mechanisms for developing T2DM are still unknown; however, there are several important risk factors. T2DM is a multifactorial disease occurring due to many factors such as obesity, hyperlipidemia, hypertension, unhealthy or sedentary lifestyle, stress, aging, ethnicity, family history of DM, and high blood glucose during pregnancy.

DM increases the risk of early death. In 2021, the International Diabetes Foundation (IDF) estimated 6.7 million death among adults worldwide because of DM and its complications [2]. Furthermore, DM is the 7th primary reason for death worldwide by 2030 [3]. Figure 1-1 exhibits global and United Arab Emirates (UAE) diabetes prevalence.

Therefore, early detection and appropriate treatment are essential to prevent disability and death. Consequently, detecting early biomarkers coupled with inhibited disease progress is imperative.

Biomarkers are classified into two types: traditional or novel biomarkers. These two types of biomarkers are distinguished based on their categories (1) clinical biomarkers, e.g., age, gender, race, and family history, (2) biochemical biomarkers, e.g., Glycemia, and (3) molecular biomarkers, e.g., deoxyribonucleic acid (DNA) based, proteomics and metabolomics. This study focuses on biomarker discovery in a metabolomics context. Several studies showed a significant relationship between metabolic abnormalities and T2DM [4-7]. The metabolic phenotype of dysregulation can be symptomatic of an aberrant biochemical or physical state.

DM is considered a major risk factor for diabetic kidney disease (DKD). DKD is a chronic condition with unknown etiology.

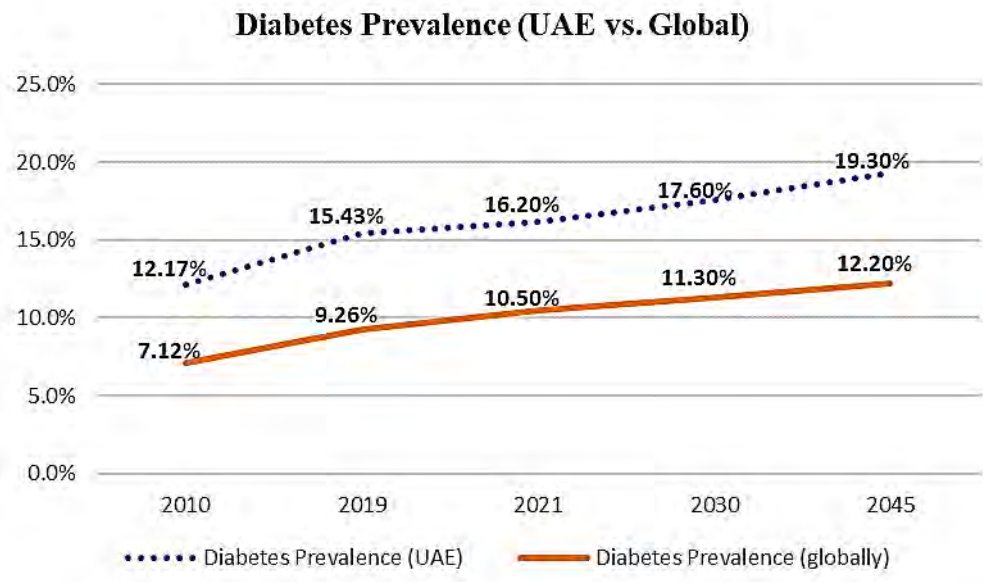


Figure 1-1: Diabetes prevalence (UAE vs. Global), related- health and economic issues [2].

1.2 Diabetic Kidney Disease

Chronic kidney disease (CKD) has become a primary worldwide health concern due to the high mortality rate [8, 9]. Individuals with CKD are five to ten times more susceptible to premature death than to progress to end-stage renal disease (ESRD) [10]. Individuals with ESRD will routinely undergo hemodialysis to compensate for the failing kidney function. DKD develops in almost 40% of patients who have diabetes

and is the leading cause of CKD worldwide [11]. DM is a leading cause of ESRD [12, 13]. Inversely, the renal function progressive decline and CKD-related sequelae also disturb glucose metabolism [14]. This association has been of long-standing interest. Cardiovascular mortality and progression to ESRD are the two significant unmet medical needs in patients with CKD and DM. Diabetic patients undergoing hemodialysis have a lower survival rate than non-diabetic patients with ESRD due to other renal diseases [15, 16].

Hemodialysis is a frequent procedure to compensate for the failing kidney function, resulting in a constant shift in the metabolic profile. For example, a recent study found that almost one-third of diabetic hemodialysis patients might face impulsive solutions of hyperglycemia with glycated hemoglobin (HbA1c) levels less than 6% [17]. This uncertain biological plausibility and unspecified medical consequences is a phenomenon called "Burnt-Out Diabetes" [18]. Further, several glucose-lower agents and their active metabolites are metabolized in the kidneys and emitted, requiring dosage correction or avoidance in hemodialysis patients [18]. Therefore, DKD patients under routine hemodialysis will encounter hyperglycemia and hypoglycemia via multifactorial processes relating to kidney dysfunction, the uremic environment, and hemodialysis [18-22].

DKD is a severe irreversible complication of DM that further disturbs glucose metabolism. Therefore, the quest for predictive and surrogate endpoint biomarkers for advanced DKD has received significant interest [13].

1.3 Biological Background

Molecules are two or more atoms with differing types, numbers, and chemical bonding. The molecular structure governs the physical and biochemical properties, such as folding patterns that change in response to microenvironmental cues in living systems. The study of molecular structure in applied biology is key to understanding biochemical mechanisms and vital properties. For example, the chemical and pharmaceuti¬cal properties of therapeutic medications depend on the molecular structure knowledge. Furthermore, biological molecules are classified into three main classes; carbohydrates, lipids, and proteins. These essential components of the cell are necessary to perform various biological functions.

Carbohydrates deliver energy through simple sugars such as glucose. Carbohydrates can be symbolized by the (CH2O)n formula, where n is the number of carbon atoms in the molecule. Namely, carbon to hydrogen to oxygen is 1:2:1 in carbohydrate molecules. Carbohydrates are classed into three subtypes; disaccharides, monosaccharides, and polysaccharides.

Lipids are nonpolar hydrophobic, water-fearing molecules that store energy long-term in the cell. Lipids also provide protection from the environment for plants and animals. For instance, they help maintain mammals and aquatic birds dry due to their water-resisting nature. Lipids are also an essential component of many hormones, such as steroids and plasma phospholipids membrane.

Proteins are polymers of amino acids arranged in a linear sequence that will later undergo post-translational modification and a unique folding pattern that dictates their molecular structure and function. Each cell in a living system might include thousands of various proteins, each unique structure, and function. For example, enzymes are catalysts in biochemical reactions like digestion and are usually proteins. Furthermore, nucleic acids hold the genetic blueprint and have instructions for the cell's functioning. The two categories of nucleic acids are DNA and ribonucleic acid (RNA). DNA is the genetic material observed in all living organisms, varying from single-celled bacteria to multicellular mammals. DNA, RNA, and polynucleotide make up monomers known as nucleotides. Each nucleotide comprises a nitrogenous base, a pentose sugar, and a phosphate group. Specific DNA sequences that translate particular proteins are known as genes and are parts of heredity.

The cell is an essential component of all living things and examines biological mechanisms. Microorganisms such as bacteria, parasites, and yeasts may consist of as few as one cell, while a highly evolved human body contains trillion cells. All cells are bound by a plasma membrane and filled with a cytoplasm that contains DNA and Ribosomes for protein synthesis. For example, the human red blood cell. A group of cells combines the tissue such as human skin. A group of tissues makes up an organ system, such as the stomach. A group of tissues comprises the organism; each human is an organism.

In essence, to conduct any given biological function in the cell in response to environmental stimuli, DNA will first be transcribed into mRNA called messenger,

followed by a translation into protein in the ribosome, that will undergo further modifications to conduct the biological function and resulting in metabolites, the aftermath story. The quest to explore the interplay between a disease genotype, a unique sequence of DNA inherited for a particular gene, and its phenotype results from the interaction between genotype and environmental factors. The multi-OMICS approach can solve nature vs. nurture.

1.4 OMICs, The Interpreting Language of Molecular Life

The first bacterial genome sequencing was back then in 1995 [23]. Subsequently, the Human Genome Project declared its first version in 2001 [24, 25]. Understanding the molecular system of living organisms has led to advancements in technological techniques to measure the function of critical biomolecules in living organisms, namely: RNA, DNA, proteins, and small molecules of diverse nature. The analysis of such elements led to the growth of the research fields known as Omics [26, 27].

Omics has become the new slogan of molecular biology. In recent years, the utility of -Omics technologies, such as genomics, proteomics, metabolomics, and transcriptomics [7], has delivered new perceptions of well-being. For example, Omics enhances monitoring disease evolution, dietary interventions, and drug toxicities by revealing the triggers of several diseases and detecting promising links between apparently different conditions [28]. The terms Omics is a derivation of the suffix -ome, which has been added to past existing biological terms like genomics, proteomics, transcriptomics, and metabolomics. Omics seek to detect the whole set of biomolecules confined in a biological fluid, cell, tissue, or organism, which creates a massive volume of data explored by biostatistics and bioinformatics methods [29]. Figure 1-2 indicates how Omics technologies are associated and their biomedical studies roles.

Omics platforms have promising avenues for biomarker discovery, identifying signals molecules related to cell death, cell growth, cellular metabolism, and early discovery of disease [30]. Genomics/transcriptomics allows evaluating potential information, proteomics evaluating executed plans, and metabolomics reveal the outcomes following these plans' execution [31].

Interactions between multiple genes and environmental factors triggered complex diseases such as cancer, DM, cardiovascular disease, schizophrenia. Therefore,

discovering the metabolomes or metabolite profiles of such conditions has earned significant consideration in the area of genomics and big data[28].

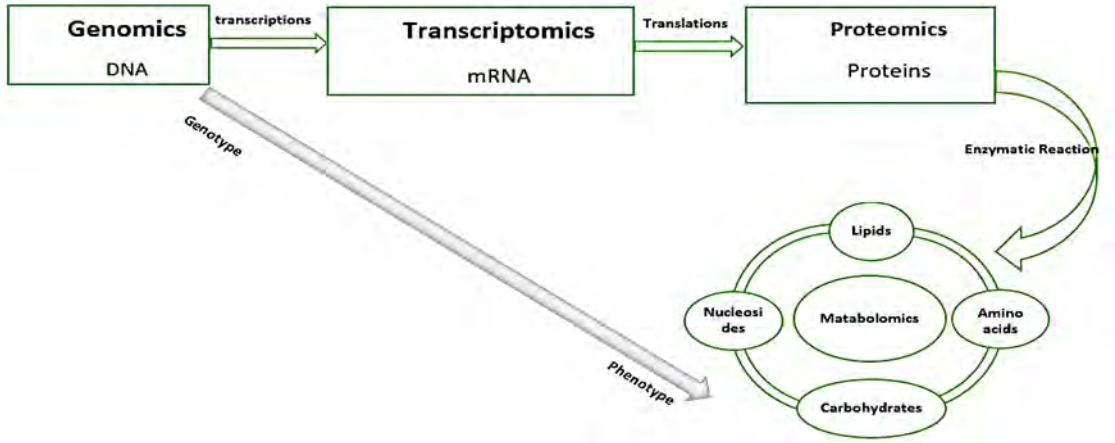


Figure 1-2: The correlation between leading omics technologies. Adapted from [35].

1.5 Metabolomics

Living organisms, such as humans, are very well-ordered systems. Several studies showed that everyone has a metabolic pattern displayed in their biological fluids' genetic makeup. Roger Williams was the first who present this theory in the late 1940s [32]. Studies described the human metabolome as the qualitative and quantitative set of all small molecules, called metabolites. They present in a cell as contributors to general metabolic reactions and are necessary for growth, maintenance, and regular function [33][34]. A unique metabolic fingerprint for each cell and tissue type can reveal organ or tissue-specific information. Biospecimens or biosamples utilized for metabolomics analysis comprise but are not limited to plasma, blood serum, saliva, urine, muscle, feces, sweat, exhaled breath, amniotic fluid, and gastrointestinal fluid [35].

Metabolites have two main types: primary/central metabolites and secondary/ specialized metabolites. The former represents the compounds directly involved in the metabolic pathways of an organism's growth, development, and reproduction. The latter also symbolizes the organic compounds produced by various organisms indirectly involved in the organism's growth, development, or reproduction. Primary metabolites contain enzymes, proteins, lipids, carbohydrates, vitamins, ethanol, lactic acid, butanol, etc., that make the organism's structural and physiological organization. Secondary metabolites contain steroids, phenolics, essential oils, pigments, alkaloids, antibiotics, etc.

Human beings are 99.9 percent similar in genetic makeup [36]. The 0.1 percent differences can deliver significant signs about the origins of diseases. Various diseases result in variations in biofluids' metabolite profiles before clinical symptoms. The metabolic signature of diseases explains the disease's pathophysiological mechanisms and proposes new drug targets. Metabolomics has demonstrated predictive, diagnostic, and prognostic abilities that have facilitated the analysis of factors affecting chronic diseases' onset and progression.

The terms metabolic profiling and metabolomics are used interchangeably [37]. The motto metabolome initially seemed in the literature in 1998. Oliver et al. [38] evaluated the change in metabolites' relative concentrations due to the deletion or overexpression. In the 1990s, researchers defined metabolomics as the methodical experiment of the distinctive chemical fingerprints that specific cellular processes leave after, precisely, studying their small metabolite profiles [39]. Metabolomics identifies and determines the collection of metabolites or specific metabolites in specimens (cells, biofluids, tissues, or organisms) in normal situations compared to progressive alterations triggered by diseases, drug treatment, nutrition, or environmental influences and genetic effects.

Chronic diseases occur from the impact of multi factors, such as genetics, lifestyle, and environment. Comparing metabolite concentration levels in phenotypically recognized populations, e.g., diseased and control subjects, might support identifying pathways and biological activities linked with a specific disease. Hence, effective computational techniques have become essential to decoding many changes' effects.

Horning et al. [41] introduced early metabolism studies. Toward the end of the 1990s, several omics acronyms were revealed. The terms metabolome, metabolomics, and metabonomic were proposed. Van der Greef et al. [42] review and examine the relationship of chemometrics and metabolomics and a timeline of metabolomics' development.

A metabolomics experiment involves targeted metabolomics, and untargeted metabolomics approaches. The targeted metabolomics approach defines a quantitative analysis where metabolites concentrations are predefined and determined [43]. An untargeted metabolomics approach is mainly the global profiling of the feasible metabolites from different biological specimens [44]. Hence, a targeted metabolomics

analysis requires considerable prior knowledge, and the completion of the experiment depends on the research hypothesis's strength. Furthermore, in targeted metabolomics, the classification of the metabolite or metabolite class of interest is identified. In contrast, for untargeted metabolomics experiments, metabolite identification is employed.

The overall number of distinct metabolites in a specific organism known as the metabolome is unknown [37]. The Estimations of metabolites based on identified pathways vary from hundreds to thousands. In the Human Metabolome Database (HMDB) metabolome database, more than 217920 annotated metabolites entries are registered [45]. However, there are free metabolomics resources, databases and libraries, such as HMDB [45], Kyoto Encyclopedia of Genes and Genomes (KEGG) [46], PubChem [47], Metlin [48], MassBank [49], LIPID MAPS [50], Chemical Entities of Biological Interest (ChEBI) [51], BioMagResBank [52] and the Small Molecule Pathway Database (SMPDB) [53].

The following section highlights the general methodological steps implemented in metabolomics experiments.

1.6 Metabolomics Experiments

The metabolomics experiment is a set of chronological steps highlighting targeted and untargeted metabolomics analyses. The metabolomics experiment is conducted as follows: sampling, sample preparation, instrumental analysis, data processing, and interpretation [54-57]. Figure 1-3 exhibits the methodological step of metabolomics experiments.

The crucial first step is identifying the research problem statement clearly and precisely. This step will define how the experiment will be designed and conducted [43]. The type of metabolomics approach should be identified in this step, i.e., targeted vs. untargeted. Next, the collected specimens, such as biofluids (urine, serum, plasma, saliva, cerebrospinal fluid (CSF)), tissues, cells, organisms, and sample size, should be identified. The sample's choice depends on the research question, i.e., biofluids used to detect biomarkers, while tissues and cells examine mechanisms related to pathophysiological processes. Also, experimental conditions, frequency of sample collection, metabolic quenching to interrupt enzymatic activity should be apparent. The immediate freezing of samples using dry ice or liquid nitrogen at −80 °C conditions is

preferred for long-term biological fluids storage [58]. This stage also should define the analytical platforms and sample preparation strategies for the experiment. High-throughput quantitative technological platforms have allowed fast and increasingly expansive data acquisition with samples as small as single cells; however, considerable hurdles remain [59].

Nuclear magnetic resonance spectroscopy (NMR) and mass spectrometry (MS) are the most popular analytical instruments for metabolomics sampling [60][40]. However, the data processing techniques differ whether a targeted or untargeted metabolomics experiment is conducted [61]. Metabolomic data are complicated and require advanced techniques to unlock the hidden biological metabolite between control and healthy samples [62].

Univariate and multivariate methods corroborate each other to attain the best results [63]. Chapter 3 explains further details about chemometrics in metabolomics.

Biological interpretation is of extreme importance for both targeted and untargeted metabolomics studies. The association of altered metabolites with respective metabolic pathways can explain the rationale and answer the initial biological question that guided the metabolomics study. Although it is uncommon to validate the results after metabolomics studies are completed, validation is believed to provide reliability for the obtained results. The initial focus of metabolomics experiments is biomarkers discovery accountable for certain diseases. The following section summarizes the main definitions of biomarkers.

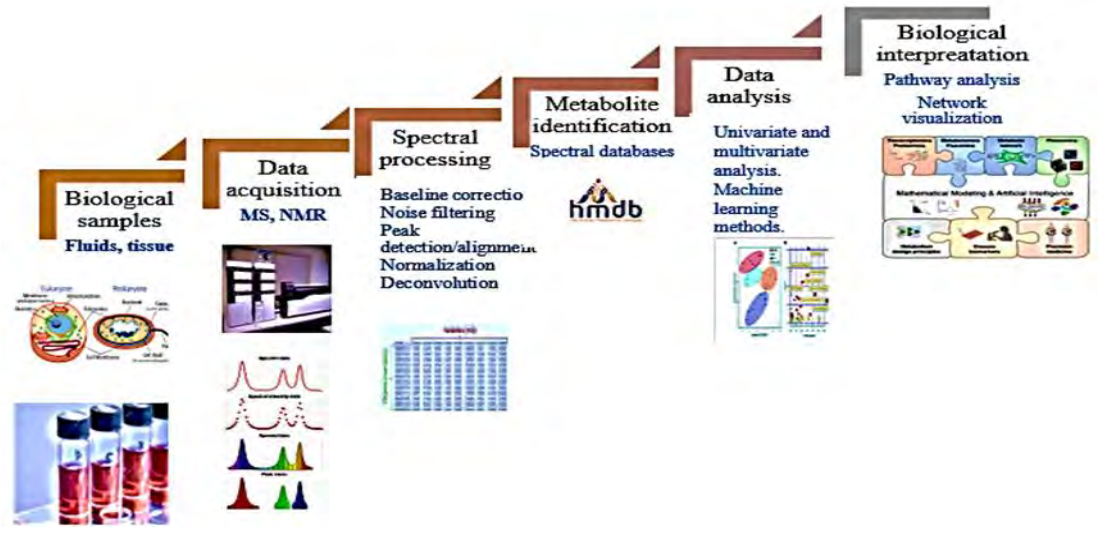


Figure 1-3: The metabolomics experiment.

1.7 Biomarker Discovery for Diabetes

The current diagnostic tools for DM, such as HbA1c, fructosamine, and glycated albumin, have limitations due to many variables such as hemoglobinopathies [64]. So, the motto one size fits all approach should not be used to diagnose or manage DM. Therefore, a vital need exists to identify more sensitive and precise biomarkers capable of predicting progression to dysglycemic states at the earliest point when the β-cell function is still relatively optimal and may be more responsive to lifestyle modification. Combining biomarkers in a clinical setting may provide better sensitivity and specificity in predicting pre-diabetes and diabetes [65, 66]. In addition, biomarkers offer the ability to identify people with subclinical disease before the development of overt clinical disorders [67]. They enable preventive measures to be applied at the subclinical stage and the responses to prophylactic or therapeutic measures to be monitored.

To date, it is unclear whether the observed metabolic changes are a consequence of high glucose levels and therefore of the diseases T1DM and T2DM or if the metabolic changes are causative and lead to the development of T1DM and T2DM [6]. We only observed the co-occurrence of both events: diabetes and metabolic changes. It seems plausible that both events occur hand in hand, especially in T2DM; thus, metabolic changes arise before or with the development of T2DM.

Metabolomics is widely employed in discovering biomarkers for disease diagnosis, prognosis, and risk prediction [68]. Diagnostic biomarkers determine the incidence and type of DM. Prognostic biomarkers deliver DM outcomes in patients to enable DM diagnosis noninvasively. Finally, predictive biomarkers support the optimization of therapy decision-making by offering information on the possibility of a reaction to a provided treatment [69, 70].

Novel biomarkers, also called molecular markers or signature molecules, are quantifiable biological molecules found in biofluids, tissues, or cells. These biomarkers can be a sign to identify, observe, or expect the risk of disease. In addition, the biomarker helps find how the body reacts to a particular medication, i.e., monitoring therapeutic measures [71]. The advancement in metabolomics technology can spot the light on unexplored territories and detect biomarkers beyond the traditional areas of urine, plasma, and serum studies [71].

Breast Cancer Type 1 (BRCA1) mutations are genetic biomarkers accountable for a substantial number of genetic predispositions to breast and ovarian cancer risk [72]. Likewise, blood glucose is a typical biomarker for monitoring diabetes or prediabetes [71]. HbA1C reveals hyperglycemia or average plasma glucose concentration over the prior two to three months of the test. Therefore HbA1C is a leading functional biomarker utilized for long-term glycemic monitoring to diagnose prediabetes and diabetes. [73]. HbA1c may also act as a biomarker that alarms for a risk factor for different diseases, such as a risk marker for diabetic retinopathy, nephropathy, and other vascular complications [71].

The risk factor may be defined as increasing the risk of disease. Risk factors can be categorized as an unmodifiable biomarker, e.g., gender, age, or modifiable biomarker such as LDL cholesterol as a risk factor for atherosclerosis [74] and smoking as a risk factor for lung cancer [75].

In the context of DM, due to the long-lasting asymptomatic clinical manifestation of DM, it is of most importance among DM researchers to discover and develop practical biomarkers with high specificity and sensitivity for the diagnosis, prognosis, and clinical control of DM. Thus, several studies found novel molecular biomarkers associated with DM.

Diabetes-related biomarkers are mainly categorized into conventional and novel biomarkers [76]. HbA1c is an example of traditional biomarkers well-defined in research and clinical medicine; however, moderate sensitivity and specificity of such biomarkers and their inaccuracy in certain clinical conditions are considered limitations. Therefore, novel biomarkers with more sensitive and accurate capabilities will boost predicting progression to dysglycemic states at the earliest time point when the β-cell function is still relatively more optimum and might be more reactive to lifestyle change [77].

1.8 Systems Engineering, Big data, and Healthcare: A Prominent Union

Systems Engineering is an established body of knowledge used for complex systems in different domains, including healthcare. Healthcare systems worldwide have incredible challenges because of the aging population, related diseases, the ever-increasing technologies use, and sedentary lifestyle. As a result, health outcomes improvement while regulating costs is a stumbling block. In this context, big data support the

healthcare sector in meeting these aspirations in distinctive approaches. Big data's promise in healthcare depends on detecting patterns and turning high volumes of data into actionable knowledge for precision medicine and decision-makers.

Big data is an integral part of the healthcare sector [79]. Medical data is produced massively and productively that requires very efficient tools to manage, store and analyze the data. Moreover, data in healthcare involves heterogeneous, insufficient, and inaccurate observations such as biological and clinical data. Therefore, various sources are used to generate high throughput profiling of such biological and clinical data cost-effectively, such as mobile phones, sensors devices, electronic health records (EHR), patients, hospitals and clinics, researchers, and other organizations. For instance, gene expression measurements using microarrays or RNA sequencing in transcriptomics and the NMR and MS platforms for proteomics, metabolomics, and lipidomic.

Big data tools available in modern software systems empower remarkable research opportunities and innovation in the healthcare domain. New emerging and interrelated paradigms such as Informatics & Data-Driven Medicine [80], eHealth [81] and mHealth [82], and Digital Health [83] are booming and attaining recognition in healthcare specialists and patients.

Big Data Analytics has emerged to perform descriptive and predictive analyses of such massive data (Figure 1-4). Databases store big healthcare data produced from several resources. Big data analytics platforms process the data for a better decision-making process. Descriptive analytics describes what happened. Diagnostics analytics answers why did it happen. Predictive analytics gives what will happen. Finally, prescriptive analytics recommends actions to affect desirable outcomes (make it happen).

Big Data Analytics is essential and popular in bioinformatics research as the human genome's size can reach 200 GB [84]. Therefore, bioinformatics researchers should develop high computational power algorithms and parallel programs.

Bioinformatics is a new, developing area that employs computational methods to solve biological questions. To answer these questions, investigators rigorously take advantage of large, complex data sets, both public and private, to reach valid, biological conclusions [85]. New research areas such as computational genomics and proteomics have arisen that target the identification of genes and their products. World Health Organization (WHO) defines genomics as studying genes and their functions and

associated techniques [86]. Together, extensive human genomic data and the innovation of analysis methods may empower more effective clinical diagnoses and new therapies. Public and private efforts were accomplished to create a comprehensive set of genes and proteins for the human genome. Therefore, the Human Genome Project generated a massive volume of genome data, an international and collaborative research program [85]. Initiatives established several projects such as the Cancer Genome Atlas (TCGA) Research [87] and the Encyclopedia of DNA Elements (ENCODE) project [88]. The most recent initiative is the Precision Medicine Initiative launched in 2015 by president Obama [89]. Precision Medicine aims to study the combined genotypes and phenotypes of at least one million volunteers that consider individual variability in genes, environment, and lifestyle for each person [89]. Precision medicine is promising for improving many aspects of health. For instance, in [90], a big-data-centered method for personalized medicine has been proposed. Moreover, by examining cardiovascular data, the American Heart Association offered a future digital ecosystem for cardiovascular disease and stroke [91].

Efficient management, analysis, and interpretation of big data can open new avenues for modern healthcare. Therefore, several healthcare industries take necessary actions to transform this potential into better services and financial benefits.

Data from DM is massive and beneficial for preventive diagnostic and prognosis of disease occurrence and outcomes. Effective management and data processing may substantially enhance human wellbeing [92]. The significant relationship between genes and diseases led to metabolomics research to unlock the hidden secret behind humans' biological processes.

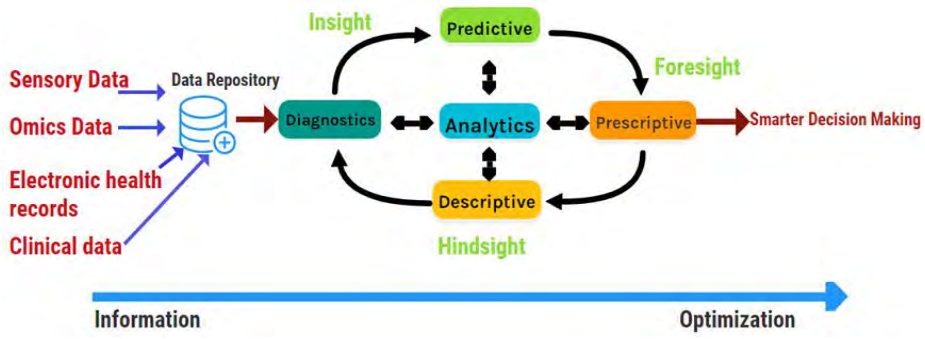


Figure 1-4: Workflow of Big data Analytics [84].

1.9 Research Motivation

DM has become an epidemic. Millions of individuals are affected worldwide, and the rates are projected to increase significantly [93]. Figure 1-1 exhibits individuals' prevalence (20-79 years) with diabetes worldwide [2, 94]. IDF reported around 537 million diabetics worldwide in 2021; however, it is estimated to reach 783 million individuals with diabetes globally in 2045 [2, 94].

The Middle East and North Africa (MENA) region ranked amongst the highest diabetes incidence rates. In 2021, one in six adults were diagnosed with diabetes, resulting in 796,000 deaths. However, the number of people with DM is projected to increase by 86% to 136 million by 2045 [2, 94].

UAE is categorized within the top countries for the prevalence of impaired glucose tolerance (IGT) in adults between 20-79 years [94]. IGT is a phase preceding diabetes when blood glucose levels are above average. Therefore, people with IGT are at high risk of acquiring T2DM diabetes, although all IGT people do not necessarily develop it.

DM prevalence in the UAE steadily increases (Figure 1-1). In 2021, the estimated prevalence of adults diagnosed with diabetes (20-79 years) was almost 16%. The expected prevalence of individuals (20-79 years) with DM in the UAE in 2045 is 19.3% [2]. Therefore, DM in the UAE is in an alarming and dangerous phase.

Moreover, the prevalence of CKD has increased considerably over the past two decades, with 13.4% of the population affected worldwide; the majority of the cases are CKD stages 3–5 [95]. In 2012, the prevalence of CKD in Abu Dhabi was unknown, but the population on dialysis was doubling approximately every five years [96]. T2DM is one of the significant risk factors for CKD among UAE nationals. It is widely accepted that the rate of CKD progression is affected by risk factors and is accelerated when multiple risk factors such as obesity, dyslipidemia, and smoking are present in an individual. However, the incidence and etiology of DKD under hemodialysis from the UAE and other middle eastern populations is unknown [97]; therefore, creating local data would support understanding the pathogenesis of DKD in the high-risk population such as UAE.

On the other hand, the Omics technologies adoption in the MENA region is nascent. Therefore, researchers should promote the concept of Omics technologies through advanced experiments, thus creating our scientific silhouette.

1.10 Research Aim and Objectives

This study aims to decode the link between metabolites and T2DM, which aids in understanding disease pathogenesis, i.e., the biological mechanism that leads to a diseased state and identifying novel biomarkers that could lead to developing more personalized nutritional and therapeutic strategies. In this study, we conducted a non-targeted metabolomics experiment using the LC-MS/MS platform available at Sharjah Institute of Medical Research (SIMR) and College of Pharmacy at the University of Sharjah to explore the profile of people with diabetes, Emirati citizens, to uncover their potential novel biomarkers. The study is twofold: (1) a comprehensive study to reveal global metabolic profiling for 50 diabetics patients versus 42 healthy and (2) a study for 11 dialysis diabetics (DD) against 25 dialysis non-diabetics patients (DND) to uncover their distinct biomarkers. Blood samples were collected from subjects in both scenarios based on clinical diabetic status and current HbA1c values. Ultimately, this study set the foundation for clinical translation by validating metabolic biomarkers associated with T2DM and other diseases. Integrating big data, computer-aided tools, and established databases and repositories helped generate a metabolic starting point for UAE studies in Omics technologies. The study outcomes are aligned with UAE goals for preventing diabetes. The potential biomarkers would be validated by conducting follow-up Omics studies in UAE. The clinical translation of novel biomarkers could expedite the treatment process and boost the healthcare system beating increasing numbers of diabetes.

This research aims to answer the research questions: what metabolites are associated with T2DM in the UAE population. In other words, are there any differences in metabolites concentration levels between group samples that reveal particular pathophysiology?. Thus, the research will examine the relationship between DM and distinct metabolites.

1.11 Research Significance

Translational medical research allows physicians to modify existing protocols to manage disease conditions and optimize patient outcomes. In addition, the development

of new technologies, tools, and drug discovery reduces morbidity and mortality. For example, insulin pump therapy innovation has prolonged life and reduced illness and disability. Therefore, it is of utmost importance to promote clinical research in our area to provide more information that may eventually lead to new medical breakthroughs. The study investigates the underlooked metabolomic role and correlation with DM in the UAE population. The significance of the proposed work can be summarized in the following points:

1. This work represents the first comprehensive approach to discover metabolomic biomarkers associated with T2DM in the UAE population.
2. DKD is a complex health disorder with unknown etiology, particularly in the middle eastern population; therefore, it is imperative to unveil the biological knowledge of such diseases.
3. The discovered metabolites will be used for future clinical trials to validate our findings as early diagnostic or prognostic tools in a clinical setting.
4. This study would contribute to a global comparison of metabolomic alterations in various populations. In addition, the results will be aligned to T2DM studies worldwide as researchers desire to pinpoint a metabolite biomarkers bank that might be potential predictors of T2DM.
5. The work will be extended to incorporate other OMICs approaches such as proteomics studies.
6. The work utilizes big data tools in metabolomics experiments.
7. We anticipate that this work may contribute to new drug discovery by outlining metabolites profiles and interactions with antidiabetic medications in UAE.
8. Technically, we hope this work attracts students and junior researchers to study and develop statistical tools that are more robust and reliable.
9. The research will contribute to the United Arab Emirates' government support of the health sector and its Diabetes Initiatives. It is in line with the Ministry of Health and Prevention’s National Strategy for Diabetes and as well as with the American University of Sharjah's strategy of taking a leadership role in contributing to the UAE and GCC challenges.
10. We hope to establish a metabolomics bank to facilitate collaboration and data access among other scientists in the UAE and the GCC region.

1.12 Execution Phases

The research studied the metabolomic profiles of T2DM individuals. The project consisted of the following steps.

1.12.1 Literature review

We examined the literature on analytical methods for metabolomics and metabolomic profiles concerning T2DM.

1.12.2 Collect biological Samples

Blood sample collection has been done at the University of Sharjah facilities utilizing the diabetes case.

1.12.3 Data acquisitions

Liquid Chromatography with tandem mass spectrometry (LC-MSMS) was used at the University of Sharjah by the Co-PI from the Department of Medicinal Chemistry. Then, a pre-processing step was carried out to remove low-frequency artifacts and differences between samples that are generated by experimental and instrumental variation.

1.12.4 Data processing

After the metabolite features are robustly quantified, multiple univariate and multivariate statistical methods have been used to perform the desired study analysis.

1.12.5 Model validation

Validation will be conducted through a follow-up study in the future.

1.12.6 Data interpretation

A comprehensive interpretation was performed to decode the link between metabolomics profiles and T2DM. The outcomes will facilitate understanding of the disease pathogenesis and identify novel biomarkers that could develop personalized nutritional and therapeutic strategies.

The project extended over two years, divided into simultaneous tasks. The different functions and their time duration are shown in Table 1. The distribution of work is also shown in the same table.

Table 1-1: Dissertation Execution Plan.

Task	Project Team				2020/2021													2021/2022											
					June	July	August	September	October	November	December	January	February	March	April	May	June	July	August	September	October	November	December	January	February	March	April	May
	PI	Co-PI1	Co-PI2	GS	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Project Planning	X	X	X	X	X	X	X																					
Literature Review	X	X	X	X	X	X	X	X	X	X																		
Design of the Research Methodology	X	X	X	X				X	X	X																		
Samples Collection			X						X	X	X	X	X	X	X	X												
Data Acquisition									X	X	X	X	X	X	X	X	X											
Data Processing	X	X	X	X										X	X	X	X	X	X	X								
Data Interpretation	X	X	X	X															X	X	X	X	X	X				
Interim Report	X	X													X	X												
Paper writeup and submission	X	X	X	X																			X	X	X	X	X	X
Final Report	X	X	X	X																						X	X	X

1.13 Dissertation Structure

The dissertation comprises eight main chapters. The first chapter introduces a background about the topic, research motivation, and aims and objectives. The second chapter includes the main profiling techniques deployed in metabolomics. The third chapter outlines the general data analysis techniques in metabolomics. The fourth chapter covers the literature review conducted in metabolomics. Chapter five includes the methodology. The sixth and seventh chapters exhibit a complete description of the first and second analyses. Finally, chapter eight concluded the work. The next chapter discusses data processing technologies in metabolomics to decrease variations or spectra incorrect phases and reducing influences differences in biofluid salt concentrations or different dilutions, disturbing the aftermath of data analysis.

Chapter 2. Data Analysis Techniques

2.1 Introduction

The conventional methodological pipeline of an untargeted metabolomics experiment combines different steps (Figure 1-3). This pipeline starts with spectral data processing to produce metabolic information [98].

After generating the metabolites, the researcher can apply univariate and multivariate data analysis. In [99], a systematic list of the broadest and free accessible tools and software mainly employed in metabolomics has been provided. Tools were classified based on the type of analytical platforms, i.e., NMR, Liquid Chromatography with Mass Spectrometry (LC-MS), GC–MS, and the role, i.e., pre-and post-processing steps, statistical analysis, workflow, and more functions. This chapter depicts the data spectral processing techniques and different metabolomics statistical analysis tools.

Spectral data processing aims to detect and quantify the molecular features, i.e., MS (mass-to-charge ratio (m/z)) spectrum and NMR spectrum peaks. Then, arrange them in a feature quantification matrix (FQM). The FQM includes the counts of all the examined samples' metabolic features. The acquired features will be used for later statistical analysis, i.e., univariate, multivariate, or exploratory [54].

2.2 Spectral Pre-processing

Preprocessing spectral data has been considered a crucial aspect of chemometrics modeling for better data quality and interpretation [100, 101], [63]. The spectral preprocessing role eliminates noises, artifacts, and weak signals due to test environments and flaws of raw data components. Different approaches and algorithms exist and have been applied for proper preprocessing. However, several factors identify selecting a pretreatment technique, such as the chosen data analysis methods, the data set structure, and the biological research question [100]. Thus, it is debatable how the executing order should be, which is sometimes more governed by practical considerations than optimal statistical analysis [63]. The most common NMR- and MS-based spectra techniques are binning, spectral alignment, baseline correlation, normalization, and scaling.

2.2.1 Binning

Binning or bucketing technique decreases the number of variables and improves data analysis. It requires splitting the NMR spectrum into small areas, usually spanning

0.04–0.05 ppm, which are adequately large to comprise one or more NMR peaks. Binning is similar to the histogram procedure, and all amounts inside each bin are added up to produce spectra with fewer variables [63]. The area under the curve (AUC) defines each bin's intensity. Several available binning techniques include equidistant (equal size) binning, Gaussian binning, dynamic adaptive binning, adaptive-intelligent binning, and more. Several non-binning methods, such as spectral deconvolution curve-fitting, direct peak fitting, and peak alignment, have been established to avoid binning's downsides. However, these approaches are best for biofluids such as plasma, serum, saliva, CSF. The primary rationale for applying binning is the excessive number of variables to manage computer memory problems and its ability to adjust small peak shifts. Later, multivariate statistical analysis is performed on the extracted bin intensities. Then, the most important peaks or bins are allocated to particular metabolites.

2.2.2 Spectral alignment

Spectral alignment is a method that iteratively moves peak positions in several spectra so that the peaks related to the same compounds can be overlaid or aligned. Peaks or unevenly altered signals across different spectra will not be matched appropriately. Thus, successive scaling steps, binning steps, and multivariate analysis of the binned or scaled intensities will be compromised [63]. In addition, minor variations can adjust multivariate statistical analysis and abstruse the biomarkers discovery or the form of metabolic profiles. Hence, it is required to use alignment algorithms as a preprocessing step to improve local signal shifts. Spectral alignment methods are grouped into segmenting strategies and warping. One of the most common NMR alignment methods for chromatographic data is correlation-optimized warping (COW) [102]. COW method employs two factors, flexibility and section length, to monitor how spectra can be warped for a reference spectrum. The section length is used to split the spectra into sections that can be stretched and compressed as much as the flexibility parameter allows by moving the sections' endpoints. Another method that employs Fast Fourier Transforms is the ice shift procedure [103]. However, some of the limitations of alignment methods may bias signal areas and compromise metabolite quantification accuracy. Therefore, absolute quantifying based on raw data is recommended. Software packages, for instance, Chenomx, offer signal shifts in individual NMR spectra and generate reliable metabolite quantification without the necessity for signal alignment.

2.2.3 Baseline correction

Raw data from different measurements usually have an unnecessary linear or non-linear addition to the spectra. Many statistical analysis techniques cannot separate the noise baseline and actual signals. Thus, they will be disturbed when the baseline is not entirely flat and can overfit the data [63], [104]. These distortions are corrected for NMR spectra data processing because they offset the intensity values and inaccuracy in peak assignment and quantification. Adjusting such errors is crucial in metabolomics, including many small but statistically significant peaks sensitive to baseline distortions. Inaccurate quantification of later peaks could lead to failures in discovering essential metabolites or detecting potential biomarkers [104]. Therefore, baseline correction is used to remove these variations between samples. It can be achieved before FT on raw data or the NMR spectrum. There are several baseline correction techniques available. The most common category is the family of polynomial fitting baselines, e.g., Lieber and Mahadevan-Jansen's iterative polynomial fitting. However, limiting the form of the baseline to a polynomial is not always the best option. Approaches like robust baseline estimation (RBE) and asymmetric least squares (ALS) use further constraints on the baseline's shape that are often more reasonable. [63] described a method for selecting baseline algorithms and their factors.

2.2.4 Normalization

Sample normalization is an essential part of the overall metabolomic profiling workflow for quantitative metabolomics. Different unwanted signal variations in metabolomics data negatively alter metabolic profiling accuracy. Samples containing metabolites can vary substantially from one sample to another caused by the variation of dilution factors for various samples. Changes in intensity for MS spectra may occur due to different quantities of metabolites reaching the detectors. For instance, urine samples may have different metabolite concentrations due to the solvent's water variations. Thus, the calculated metabolite concentrations will indicate dilution instead of the changes in the metabolic response. Therefore, reducing or eliminating the total sample amount variation on individual metabolites' quantification. Accordingly, various normalization methods can be applied to address the problems mentioned above after peak alignments, identifications, or binning, and the determination of their respective intensities. The two main applied ways are stable endogenous metabolites like creatinine in urine and the total spectral area, i.e., AUC. When employing

univariate analysis, it is noteworthy that it is unnecessary to apply variable normalization because each metabolic feature is assessed individually. Conversely, normalization in multivariate analysis is extremely valuable and relies on the research question.

2.2.5 Scaling

Scaling refers to a statistical technique that improves the normality distribution of data or lowers the dispersed values. It employs a mathematical operation on the spectra intensities or concentrations. Metabolite concentrations can vary over numerous orders of magnitude. This fact can make a small number of metabolites dominate the outcomes from multivariate statistical analyses. To avert this bias, scaling metabolite intensities is necessary before further study. Centering adjusts the differences between low-concentration and high-concentration metabolites. Centering scales each value to fluctuate around zero, where zero is the mean metabolite level. Depending on the experimenting nature, a range of statistical and data mining methods can be employed on metabolomic data. In the next section, both univariate and multivariate statistics are illustrated.

2.3 Statistical Analysis

After the metabolite features are robustly acquired and quantified, researchers can apply statistical methods and data mining approaches to extract relevant information from metabolomic data. There are two main data analysis methods, univariate and multivariate methods. The latter techniques also are identified as chemometric methods. Here, we describe the most applied metabolomic features and the most used chemometric methods.

2.3.1 Metabolomics features

The preprocessed metabolomics data, both MS and NMR, is typically organized into an FQM. In this matrix, rows relate to the samples, and columns relate to the obtained metabolomic features. The metabolomic feature is almost associated with the concentration of a metabolite. Data analysis techniques can then be applied using these metabolomic features as input.

2.3.2 Univariate analysis methods

The univariate analysis provides an initial summary of the data features possibly significant in differentiating against the study conditions. At This Point, only one

metabolomic feature is analyzed at a time. These techniques are standard easy to apply and interpret. The main limitation is that they do not consider the relationship between distinct metabolic features. Furthermore, the non-considerable impact of possible confounding variables like diet, gender, or body mass index (BMI), increases the likelihood of getting false-positive or false-negative outcomes. Nevertheless, univariate analysis methods can analyze metabolomic data. For instance, parametric tests such as ANOVA and Student's t-test are typically employed when evaluating variations between two or more groups, assuming that normality assumptions are validated.

2.3.3 Multivariate analysis methods

Contrary to univariate approaches, multivariate analysis methods consider each of the metabolomic features at once to discover associations between them. The main two multivariate methods can be categorized into supervised and unsupervised methods. Examples of such methods include multivariate regression analysis, multivariate ANOVA, principal component analysis, factor analysis, and partial least square discriminant analysis. Supervised methods use the sample labels to distinguish the features or feature combinations related to a phenotype of interest. They are also the base for developing prediction models. Most multivariate analysis applications in metabolomics apply principal component analysis (PCA) for data exploration. Then, OPLS-DA or OPLS for regression, class discrimination, or biomarker discovery.

2.3.4 Unsupervised methods

Unsupervised methods summarize the complex metabolomic data. They help detect data patterns associated with biological and experimental variables. PCA is the dominant unsupervised method in metabolomic studies used for visualization and exploration. PCA is the linear transformation of the metabolic features into linear and uncorrelated (i.e., orthogonal) variables named principal components. PCA is considered a starting point of any analysis to detect trends, groups, and outliers. Also, unsupervised methods such as hierarchical clustering analysis (HCA) and self-organizing maps (SOMs) have been employed to metabolomic data. These methods are especially appropriate to uncover non-linear trends in the data that PCA does not easily cover. For example, some metabolomics studies use SOMs to visualize feature patterns and metabolic phenotypes and arrange the specified metabolites based on their similarity.

2.3.5 Supervised methods

Supervised methods identify metabolic patterns linked with the phenotypic variable of interest. These methods are also the basis for developing classifiers based on metabolomic features. Partial least squares are the most broadly employed supervised methods in metabolomics. It performs as regression analysis, i.e., the quantitative variable of interest, or as a PLS-DA binary classifier, i.e., the binary variable of interest. However, PLS's weakness is that some metabolic features that don't correlate with interest variables may manipulate the results. Therefore, orthogonal PLS (O-PLS) was developed. Support vector machines (SVMs) are supervised analysis methods to establish classifiers based on metabolomic data.

Although there is a problematic interpretation of classifiers based on SVM, they can handle non-linear relationships between the variable of interest and the metabolomic data.

To summarize, numerous data analysis methods have been offered for the spectra generated from LC-MS and NMR spectrometers. PCA and PLS methods are the dominant, as commercial and open data analysis tools widely adopt them. In contrast, the non-frequent analysis methods include HCA, batch-PLS, k-nearest neighbors (KNN), orthogonal signal correction (OSC) combined with PCA, and various neural network applications. However, some challenges arise with metabolomics data analysis. For example, the multivariate projection methods (i.e., PCA and PLS) search for the most robust data variations and dominate the first new components. Consequently, more subtle but significant differences are characterized by higher elements often overlooked by the data analyst or spread over numerous features that are not detectable to the data analyst [37]. Therefore, new data analysis methods for metabolomic are introduced, such as O-PLS [105] or statistical correlation techniques, for instance, the statistical total correlation spectroscopy (STOCSY) [106].

2.4 Pathway Analysis

Two of the main challenges in Omics data analysis are the dimensionality dilemma produced by more variables than samples and the development of algorithms that successfully integrate and analyze biological data, incorporating present and future knowledge. Pathway Analysis (PA) has developed and established a reliable answer to manage these issues.

PA, also known as functional enrichment analysis, is one of the commonly used principal tools of Omics research. PA tools analyze data obtained from high-throughput technologies, identifying potential perturbed genes in diseased samples compared to a control. In this sense, PA methods aspire to conquer the dilemma of interpreting overwhelmingly large lists of essential genes, the main output of most basic high-throughput data analysis. In addition, PA methods provide meaning to experimental high-throughput biological data, therefore, enabling interpretation and successive hypothesis generation. PA targets have been reached by combining biological knowledge from databases with statistical testing, mathematical analyses, and computational algorithms [107].

PA methods hold a wide scale of applications in physiological and biomedical research. PA aims to benefit the researchers by discovering biological themes and which biomolecules are crucial to insight into the phenomena under study. The generated clues empower the researcher to create new theories, design the following assays, and validate their outcomes. For example, PA methods have supported researchers in identifying the biological functions of potential genes selected to develop new treatments for cancer [108].

PA needs several elements to operate. At the outset, quantitative data on cell biology is produced through Omics technologies such as RNA-microarrays, LC-MSMS, and RNA sequencing. Then, a method to analyze such massive information is needed. Next, databases store the molecular biological knowledge for downstream analysis, leading PA methods to explore links between the Omics data and common biological themes. Finally, computer-aided tools are employed to achieve PA. These tools comprise statistical testing of the biological themes against the data and mathematical algorithms to obtain associations between the data and prior knowledge.

The PA workflow starts with the input phase, including choosing a PA method, analyzing Omics data, and extracting the pathway data from the database. Then, the analysis phase involves all statistical and mathematical computations accomplished by the PA method. Even though the used algorithms are varied, they share commonalities and are driven by the same method. PA computer-aided tools can be found in three different styles: stand-alone software, web-based applications, and programming packages. The first two categories are user-friendlier than packages as they require less

analytical skills or programming-related talents. R and python are usually employed to code programming packages. The main benefit of using PA programming packages is the customization ability of the analysis and the possibility of automation through scripted analysis pipelines. Choosing between different platforms relies on client skills and the cost-benefit ratio of time invested in arranging everything necessary to run the analysis. Finally, the output phase covers visualization and results from the study. The results presented a ranked list of relevant pathways, and the top pathways are often ordered based on P-value or the multiple testing corrected q-value.

Additionally, directed acyclic graphs are formats used to visualize results in which relevant categories are hierarchically ranked according to their relationship, such as within the Gene Ontology categories. Heatmap formats are similarly utilized to visualize results since pattern generation among related pathways and samples is simpler to explore in this approach. Additionally, most web-based and stand-alone software provide links to web pages in databases and other online resources for easier integration of the results.

PA methods mainly refer to over-representation analysis (ORA) or enrichment analysis. The primary premise in ORA is to understand complex biological systems. These methods reduce data complexity, enhance interpretation and insight of biological systems, and better yield hypotheses. ORA searches for keywords or descriptors of the set of molecules of interest, e.g., over-expressed molecules, concerning a background reference set, e.g., the whole genome/transcriptome/proteome/metabolome or the collection of molecules identified by the technique used [109].

In this manner, ORA methods act along the main workflow of statistically evaluating the fraction of pathway components found among a user-selected list of biological features [107]. The list fulfills typically specific criteria: log fold change, statistical significance or both, ranking and cutting off most components from an original list, for instance, all molecules tested in a metabolomics experiment [107].

Classical enrichment analyses use Fisher's exact test, but many other enrichment methods, such as hypergeometric, Kolmogorov–Smirnov, or Wilcoxon statistical tests, have originated from it. Finally, multiple testing correction is usually achieved as assessing data with several hypotheses at once (in this case, pathways) causes false

positives. The outcome from an ORA method consists of a list of the best relevant pathways, ranked based on a p-value or a multiple-hypothesis-test-corrected p-value.

However, although ORA can find meaningful insights in large biological data sets, these methods have several limitations, including (1) neglecting a large amount of basal level data because of the user-selected cut-off method. In addition, the potential significant elements close to the cut-off threshold are frequently omitted in the analysis leading to repercussions in results stability. Finally, the arbitrary selection of the cut-off thresholds generates different results as there is no general principle for establishing a cut-off threshold. And (2) ORA assesses every element in the pathway, providing them the same weight or importance, neglecting any information inherent to the interactions, such as location in the pathway and interaction between molecules. Thus, two pathways with the same molecules but different topologies would generate the same result [110]. Lastly, (3) ORA assumes that pathways are independent of each other, against acknowledging the interaction and overlapping between pathways [111].

The subsequent PA generation is pathway-topology based (PTB). The fundamental premise of PTB analysis is that interactions observed in pathway topology, annotated in databases, bear information for interpreting correlated changes between pathway components. PTB techniques are expansions of the ORA methods, they act along with the same general steps, but they add pathway topology for evaluating the statistical relevance of the pathways.

PA is a steadily growing and developing interdisciplinary research area. Most current methodologies are designed to use pathway topology information stored in different databases and all the data from Omics technologies. Yet, some challenges need to be overcome to identify relevant pathways better. The challenges can be managed in three classes: advancements in Omics technologies, annotations of databases, and PA algorithms progress and use. First, there is no consensus on a standard definition of pathways and usage of a pathway ontology. The lack of cohesiveness might be due to every single project biological focus had in its beginnings or cellular functions being very different from being confined in a specific paradigm. Again, though, the unification of similar ontologies and usage of universal languages should be the most pragmatic and feasible solution across databases and PA methodologies. Also, databases lack annotation depth and coverage.

PA methods should certainly not be black boxes from where experimental data goes in, and factual statements come out, however perhaps more as cleaners of haystacks from where we are pursuing meaningful biological needles [107].

Chapter 3. Literature Review

3.1 Metabolomics and Diabetes

The encouraging role of metabolomics in translating biomarkers to the clinic is remarkable. Therefore, steady growth in the number of yearly publications, including both "metabolomics" and "biomarker" in the past ten years (2009–2021), is observed (Figure 3-1). In 2021, rapid progress happened, with over 2000 articles on metabolomics-aided biomarker research. The following sections explain DM's developmental stages and how metabolomics stands in DM research.

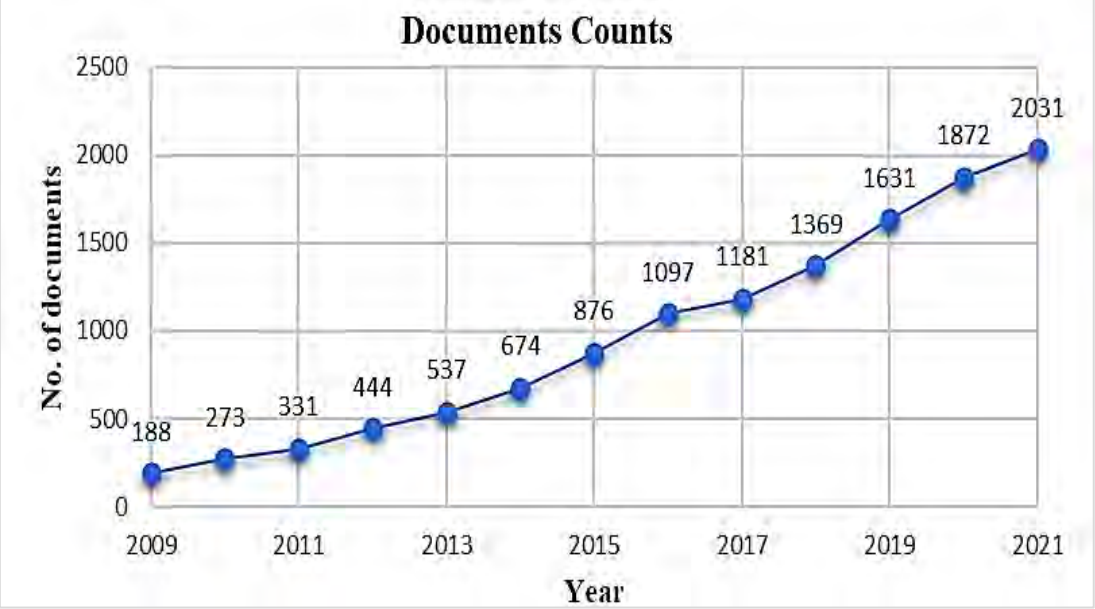


Figure 3-1: Scopus research results using "metabolomics" and "biomarker" (2009-2021).

3.2 T2DM Pathogenesis

Insulin resistance (IR) is a crucial reason for T2DM that primarily refers to the decrease in the body's biological effects on specific insulin concentrations. IR leads to impairment in glucose uptake and metabolism of the body, representing the decrease of insulin sensitivity and, therefore, the deterioration of responsiveness [112]. IR is stimulated by receptor loss, sequence mutations, mitochondrial dysfunction of skeletal muscle [113], and the actions of cytokines like free fatty acids, tumor necrosis factor, leptin, resistin, and adiponectin. Impaired islet cell function is related to islet α and β cells. The number of islet β cells is remarkably decreased in T2DM patients, and the ratio of α/β cells is elevated dramatically. In addition, the sensitivity of α cells to glucose is decreased, which increases the glucagon level and liver sugar output and finally results in the prevalence of T2DM [114]. This is the conventional hypothesis of double

hormone abnormalities [115]. IR can disturb glucose and lipid metabolism levels in the body, causing overstated free fatty acids.

Exaggerated free fatty acids stimulate the body to produce large amounts of reactive oxygen species (ROS) and reactive nitrogen species (RNS), which yield oxidative stress [116]. Oxidative stress activates the nuclear factor-kB (NF-_kB) signaling pathway by disrupting the mitochondrial structure and inducing apoptosis, which causes cellular inflammatory responses and inhibits insulin synthesis and secretion. Oxidative stress also affects physiological processes related to insulin signaling, involving the phosphorylation of insulin receptor (InsR) and insulin receptor substrate (IRS), activation of phosphatidylinositol 3-kinase (PI3K), and glucose transport of protein 4 (GLUT4) to generate IR. Also, Oxidative stress damages the anatomical structure and later yields T2DM [117]. T2DM patients are commonly obese, especially centrally obese, mainly characterized by abnormal glucose and lipid metabolism. There is a substantial negative correlation between the abdominal fat area and insulin-mediated glucose utilization. Central obesity patients have increased abdominal fat, metabolic disorders in visceral adipose tissue (VAT), and impaired hepatic glycogen production by insulin leads to IR [118]. Free fatty acids increase in the liver and muscle leads to lipid metabolites accumulation. Lipid metabolites accretion might generate dyslipidemia, impaired β cell secretion of insulin, and exaggerated fatty acids inhibit glucose from clearing, leading to T2DM [119]. The general pathogenesis of T2DM is shown in Figure 3-2 [120]. T2DM is characterized as a decrease in insulin sensitivity. The leading reasons for T2DM are obesity, oxidative stress, gene and aging, and IR. The rise of visceral fat implies increased fatty acids, increasing gluconeogenesis and glucose levels. The increased glucose level affects the compensation and decompensation of β cells, leading to IGT and eventually T2DM development. However, the metabolomics role in DM studies is promising. Detailed information is in the next section.

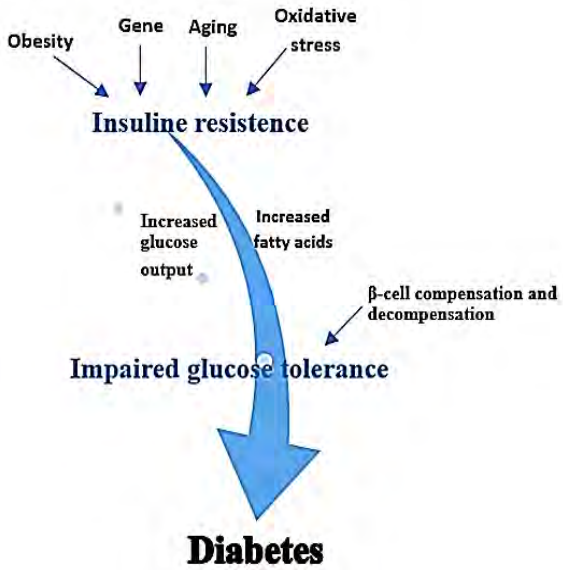


Figure 3-2: General pathogenesis of T2DM [115].

3.3 Metabolomics Footprint in T2DM Pathogenesis

Recently, many researchers have studied T2DM metabolites and revealed potential biomarkers and their metabolic pathways. These findings provide a necessary speculative foundation for T2DM successful prevention and treatment. The main factors that can normalize insulin secretion in the human body include glucose, amino acids, fatty acids, neurotransmitters, and hormones. Islet cells sustain specific homeostasis in several states by managing and integrating these regulatory factors. Conversely, T2DM can cause many metabolic abnormalities of substances in the body, such as amino acids, lipids, carbohydrates, and others. Metabolomic techniques can detect irregularities of metabolites by identifying early biomarkers that can predict the incidence and advancement of diabetes. These potential biomarkers are essential for studying the inhibition and therapy of diabetes. Many studies have revealed that perturbed metabolites significantly after the onset of T2DM are primarily amino acids, lipids, and carbohydrates. Findings have demonstrated that amino acids are potential biomarkers of T2DM and branched-chain amino acids (BCAA), including leucine, valine, and isoleucine. These are vital amino acids for human beings. Studies found BCAA closely linked to IR and diabetes [121, 122].

On the other hand, diabetes is usually associated with dyslipidemia [123], yet the exact mechanism is unclear. For instance, experiments [124, 125] showed that free fatty acids might be the leading reason for IR. A recent study [126] introduces a comprehensive review of perturbed lipids and amino acid metabolites related to T2DM. Carbohydrates

are comprised of sugars, such as fructose and glucose, formed of monosaccharides or two disaccharides. Simple carbohydrates are smoothly and rapidly consumed for energy by the body due to their simple chemical structure, commonly causing a sharper increase in blood sugar and insulin secretion from the pancreas, which can have adverse health effects. Several studies applied metabolomics to examine the pathogenesis of DM. A group of researchers [127] reviews the metabolomics application in diabetic complications. The study includes diabetic coronary artery disease, diabetic nephropathy, diabetic retinopathy, and diabetic neuropathy. The study of diabetic complications is still in the early stage in metabolomics research. Also, there are few studies on biomarkers discovery for diabetes.

Yun et al. [4] conducted a targeted metabolomics experiment to identify metabolites linked with the level of HbA1c in the serum of people with diabetes. The study identified twenty-two metabolites in the discovery set and validated sixteen in the replication set. Multivariate logistic regression analysis was performed to cluster the metabolites based on their concentration differences (low/high levels) depending on the level of HbA1c. Metabolites with high concentrations in the normal HbA1c group, such as glycine, valine, and phosphatidylcholines (PCs), could improve HbA1c levels in diabetic patients. The metabolite signatures discovered in this research give insight into the pathogenesis in HbA1c levels in T2DM.

Arneth et al. [6] explore the current published research to assess the relationship between metabolites and T1DM and T2DM. All reviewed studies signify the relationship between multiple metabolites and diabetes. For example, metabolites such as glucose, fructose, amino acids, and lipids are typically altered in individuals with T1DM and T2DM.

[128] reviewed the discovered metabolites associated with prediabetes and T2DM. The study selected 27 cross-sectional and 19 prospective publications for the systematic review. It has been found that different blood amino acid is consistently associated with the risk of developing T2DM. The blood concentrations of numerous metabolites, including hexoses, branched-chain amino acids, aromatic amino acids, phospholipids, and triglycerides, were associated with the incidence of prediabetes and T2DM.

A 6-year follow-up Chinese study [129] was conducted to pinpoint metabolites linked with an increased risk of T2DM. The serum samples of individuals were analyzed for

metabolic profiling purposes. The sample includes 197 diseased individuals with T2DM and without cardiovascular or cancer diseases before the diabetes diagnosis and 197 healthy controls matched by sex, age, and date of blood collection. The results of the study revealed 51 differential metabolites between diseased and healthy. Of these, 35 were substantially correlated with diabetes risk in the multivariate analysis. Some of these metabolites are chain amino acids (leucine, isoleucine, and valine), nonesterified fatty acids (palmitic acid, stearic acid, oleic acid, and linoleic acid), and lysophosphatidylinositol (LPI) species (16:1, 18:1, 18:2, 20:3, 20:4 and 22:6).

A multiplatform metabolomics study [130] to investigate diabetes used 40 individuals with T2DM and 60 controls (male, over 54 years) from the participants of the population-based KORA (Cooperative Health Research in the Region of Augsburg) study. The known biomarkers identified in the study include sugar metabolites (1,5-anhydroglucoitol), ketone bodies (3-hydroxybutyrate), and BCAA.

A research group [131] identified candidate biomarkers of pre-diabetes for a subcohort without T2DM of 876 S4 participated in the study. Of them, 91 developed T2DM incidences during the 7-year follow-up samples in the population-based Cooperative Health Research in the Region of Augsburg (KORA) cohort. The study indicated three significant metabolites (glycine, lysophosphatidylcholine (LPC) (18:2) and acetylcarnitine) had changed levels in IGT individuals as compared to those with normal glucose tolerance.

Research in the Region of Augsburg (KORA) [132] consisted of population-based surveys and follow-up periods in Augsburg in southern Germany. A subcohort without T2DM 876 S4 people participated in the study. Of them, 91 developed T2DM incidences during the 7-year follow-up. Hexose, phenylalanine, and diacyl-phosphatidylcholines C32:1, C36:1, C38:3, and C40:5 were significantly related to T2DM risk in a positive manner.

Another study on T2DM and impaired fasting glucose (IFG) was performed on plasma samples from a large population-based cohort of 2204 females from TwinsUK [133]. In this research, 3-methyl-2-oxovalerate was the strongest predictive biomarker, and it was confirmed in 720 plasma samples from an independent population. Also, the findings were validated in 189 twins, with urine metabolomics taken concurrently as

plasma. The results confirmed an overt role in the catabolism of branched-chain-amino-acids in T2DM and IFG.

A study [134] reported a metabolic signature shift for 24 IFG, 27 T2DM, and 60 ND. IFG and T2DM had significantly raised fructose, α-hydroxybutyrate, alanine, proline, phenylalanine, glutamine, BCAA (leucine, isoleucine,andvaline), low carbonnumberlipids (myristic, palmitic,andstearic acid), and significantly reduced pyroglutamic acid, glycerophospohlipids, and sphingomyelins compared to ND.

One of the few studies on insulin resistance (IR) in children is Mastrangelo et al. [135]. In this study, metabolites related to inflammation and central carbon metabolism, together with the contribution of the gut microbiota, were recognized as the most altered processes. Most metabolites differing between groups were lysophospholipids (15) and amino acids (17). Bile acids exhibit the most remarkable changes. Sex proved a strong influence in selecting the metabolite markers despite their prepubertal status.

Lie et al. [136] explored the mechanism of the complex disease, T2DM coronary heart disease (T2DM-CHD). The study analyzed plasma samples from 15 HC, 13 coronary heart disease (CHD) patients, 15 T2DM patients, and 28 T2DM-CHD patients. About 17 metabolic biomarkers were highly possible to be associated with T2DM-CHD. These metabolites included isoleucine, valine, isopropanol, alanine, leucine, acetate, proline, glutamine, arginine, trans-aconitate, creatine, creatinine, glucose, glycine, threonine, tyrosine, and 3-methylhistidine.

In 2016, [137] surveyed previously identified metabolic shifts in DM. BCAA, aromatic amino acids (AAAs), and acylcarnitines are strongly associated with early IR.

A meta-analysis [128] of 27 cross-sectional and 19 prospective publications reporting associations of metabolites and pre-diabetes and/or type 2 diabetes was conducted. Carbohydrate (glucose and fructose), lipid (phospholipids, sphingomyelins, and triglycerides), an amino acid (BCAA, aromatic amino acids, glycine, and glutamine) metabolites were higher in individuals with T2DM compared with control subjects. Prospective studies provided evidence that blood concentrations of several metabolites, including hexoses, BCAA, aromatic amino acids, phospholipids, and triglycerides, were associated with the incidence of pre-diabetes and type 2 diabetes.

Tam et al. [138] explored the pathogenesis and phenotype of late-onset T2DM. The study consists of a urine sample for 80 older people with late-onset T2DM and 79 older controls without T2DM. The results identified potential biomarkers; reduced levels of phenylalanine, acetylhistidine, and cyclic adenosine monophosphate (cAMP) were found in urine samples of T2DM subjects. Elevated levels of 5′-methylthioadenosine (MTA), which previously had only been implicated in an animal model of diabetes, was found in the urine of older people with T2DM.

A T2DM follow-up study [139] examined 2776 individuals from the Erasmus Rucphen Family study. Of them, 1571 healthy controls were followed up to 14-years. The results showed 24 biomarkers, i.e., high-density, low-density, and very low-density lipoprotein sub-fractions, specific triglycerides, amino acids, and small intermediate compounds predicted future T2DM.

An interesting and neglected biological sample type is earwax in [140] to detect biomarkers of diabetes. The authors studied the volatile compounds in the sample by headspace Gas Chromatography Coupled to Mass Spectrometry (GC-MS). The six most essential biomarkers were ethanol, acetone, methoxyacetone, hydroxyurea, isobutyraldehyde, and acetic acid. For example, the Methoxyacetone biomarker perfectly differentiated between T2DM and T1DM.

A prospective Swedish study [141] identified potential metabolites for T2DM future prediction. A sample of 503 case-control pairs at baseline and samples from a subset of 187 case-control pairs at ten years of follow-up were analyzed. The study identified 46 predictive plasma metabolites of T2DM. PCs containing odd-chain fatty acids (C19:1 and C17:0) and 2-hydroxyethanesulfonate were associated with the likelihood of developing T2DM.

The Singapore Chinese Health Study (SCHS), a population-based study in Singapore, investigated T2DM risk and prevalence in the Chinese population [142]. Participants involved 160 incidents and 144 prevalent cases with T2DM and 304 controls. The study recognized 37 metabolites associated with prevalent T2DM, including 7 lysophosphatidylinositol (LPIs), 18 non esterified fatty acids (NEFAs), and 12 acylcarnitines and 11 metabolites associated with incident T2DM, including 2 LPIs and 9 NEFAs. Then, LPI (16:1) and dihomo-g-linolenic acid indicated independent associations with incident T2DM and improved risk prediction considerably.

The Korean community-based cohort of the Ansan–Ansung study prospectively analyzed the associations between serum metabolites and T2DM risk [143]. A sample of 1939 participants with available metabolic profiles without DM, cardiovascular disease, or cancer at baseline was selected. In the follow-up period, the study identified 282 cases of incident T2DM. Serum levels of alanine, arginine, isoleucine, proline, tyrosine, valine, hexose, and five phosphatidylcholine diacyls were positively associated with T2DM risk. In contrast, lyso-phosphatidylcholine acyl C17:0 and C18:2 and other glycerophospholipids were negatively associated with T2DM risk.

The prospective study [144], Atherosclerosis Risk in Communities (ARIC), analyzed known metabolites using an untargeted approach in serum. A sample of 2939 participants with metabolomics data and without prevalent diabetes was selected. The study identified 245 metabolites. Seven metabolites were significantly associated with incident diabetes, including a food additive (erythritol) and compounds involved in amino acid metabolism [isoleucine, leucine, valine, asparagine, 3-(4-hydoxyphenyl)lactate] and glucose metabolism (trehalose). This study is the first to report asparagine as a protective biomarker of diabetes risk.

The investigation of T2DM in young adults was conducted using four Finnish cohorts [145]. Out of 229 metabolic measures, 113 were associated with incident T2DM in a meta-analysis of the four cohorts. Branched-chain and aromatic amino acids and triacylglycerol within VLDL particles and linoleic n-6 fatty acid and non-esterified cholesterol in large HDL particles are amongst the strongest biomarkers of diabetes risk.

T2DM is a multifactorial disease; therefore, since obesity is associated with an increased risk of IR and T2DM, a study [146] aimed to depict the serum metabolomic fingerprint and multi-metabolite signatures associated with IR and T2DM. a sample of 30 adults of normal weight, 26 obese adults, and 16 adults newly diagnosed with T2DM were chosen. The identified IR potential biomarkers include amino acids (Asn, Gln, and His), methionine (Met) sulfoxide, 2-methyl-3-hydroxy-5-formylpyridine-4-carboxylate, serotonin, L-2-amino-3- oxobutanoic acid, and 4,6-dihydroxyquinoline. However, T2DM was associated with dysregulation of 42 metabolites, including amino acids, amino acids metabolites, and dipeptides.

Researchers conducted a non-targeted urine metabolomics experiment [147] to understand better the role of the urine metabolome in predicting the risk of T2DM. Urine samples from two community cohorts of 1424 adults were analyzed by ultra-performance liquid chromatography/mass spectrometry (UPLC-MS). The study proposed 3-hydroxyundecanoyl-carnitine as a potential biomarker for T2DM.

Satheesh et al. [148] examined previous targeted metabolomics-based prospective studies on potential biomarkers for T2DM. The analysis revealed that many studies showed a direct association of BCAA and an inverse association of glycine with T2DM.

A comprehensive systematic review [5] revealed potential biomarkers for T2DM patients. Amino acids such as BCAA and AAAs had positive associations with T2DM. In particular, prospective studies indicated that isoleucine, leucine, valine, tyrosine, phenylalanine, glutamate, alanine, valerylcarnitine (C5), palmitoylcarnitine (C16), palmitic acid, and linoleic acid were associated with higher T2DM risk. In contrast, serine, glutamine, and lysophosphatidylcholine C18:2 decreased the risk of T2DM.

Studies about the metabolomic profile of T2DM from the Middle Eastern populations are still in their early stages. A study [149] on UAE T2DM nationals revealed significant differences in many metabolites, including BCAA, trimethylamine N-oxide, β-hydroxybutyrate, trimethyl uric acid, and alanine. A targeted MS approach showed substantial differences in lysophosphatidylcholines, phosphatidylcholines, acylcarnitine, amino acids, and sphingomyelins; Lyso.PC.a.C18.0, PC.ae.C34.2, C3.DC..C4.OH, glutamine and SM.C16.1 are the most significant metabolites [149].

Table 3-1 summarizes common diabetes-related potential metabolites. Collectively, we can sort the panel of potential discovered metabolic signatures into different pathways: (1) carbohydrate metabolism, (2) amino acid and derivative metabolism, (3) Glycolysis and TCA Cycle, and (4) lipid metabolism [150, 151].

To summarize, human metabolomics studies are susceptible to clinical confounding factors that may lead to false conclusions, as equivalent studies with different results have shown. Therefore, applying metabolomic profiling to large population-based epidemiological cohorts will enable more robust findings and reproducible results translated into real clinical markers [68]. The importance of metabolites as potential biomarkers is several. Metabolites changes earlier and more significant compared to genes or proteins, and those changes can be measured in absolute terms, while genes

and proteins demonstrate changes in activity in a different manner than those in the concentrations; metabolites can be allocated in biochemical pathways, and therefore, their changes can be biologically explained in most cases, strengthening their value [68]. The limitation of biomarkers discovery studies is that small sample sets influence the selection of controls and might have confounding factors that strongly affect the test outcome. This challenge explains why the studies used for the same disease, in the same sample type, and with the same instrumental technique often lead to different findings. Examining common metabolite biomarkers in various studies and a biochemical relationship with the disease will identify those with higher potential [68].

Challenges that require more recognition of the scientific community in metabolomics disease studies are research design, sample collection, quality, data quality assurance, reliable means of data analysis and model validation, and confirmation of metabolite biomarkers [152], [153]. Many efforts were proposed to construct metabolomics format standards as a common language between researchers, i.e., creating standard practices to boost the efficiency, validity, and understanding of metabolomics data [152], [154]. It is believed that these standards can ensure regularizing the structure and reporting of data. Also, they ease the way of data distribution, exchange, and reanalysis. In this regard, a recent success [155] has been made as a collaboration between researchers of the Metabolomics Standards Initiative, Proteomics Standards Initiative, and the Metabolomics Society. The "mzTab-M" data standard tool is developed to provide a common output format from analytical platforms using MS on small molecules. The tool can represent final quantification values from analyses and the evidence trail in terms of features measured directly from MS (e.g., LC-MS, GC-MS, DIMS, etc.) and distinct types of methods employed to recognize molecules. It also allows the removal of vagueness in identifying molecules that enable clear communication to readers of the files.

However, many identified biomarkers from cross-sectional epidemiological studies are inadequately potent to provide a clinically robust diagnosis of diabetes. The study's limitations are commonly acknowledged as single and small cohort studies, prompting the need for independent validation in well-designed, largescale studies in the future. There is a great potential for many better biomarkers to be discovered, which is a highly dynamic field of research in metabolomics [156].

Some considerations should be applied to transfer the discovered biomarkers into clinical settings [118]. First, the validation of the results should be conducted in different stages, from analytical validation to validation in independent sets of samples, employing thousands of samples from various sources. The utilization of metabolomic experiments to significant sample-based epidemiological cohorts will lead to improved and robust conclusions—high sensitivity, specificity, and reproducibility rates transformed into real clinical indicators.

However, metabolomic profiles are significantly affected by environmental and demographic factors, i.e., age, gender, blood pressure, BMI, and smoking, that alter the range of natural values and raise the possibility of false biomarker discovery [157]. Therefore, it is necessary to scrutinize the study design features that help boost the utility of metabolomics data across demographic groups. Thus, the following section summarizes the explored relationship between demographic risk factors and metabolites.

Table 3-1: Survey of discovered potential diabetes-related metabolites.

Ref	Disease / Treatment	Training Size	Biomarkers	Biological Matrix	Analytical Platform
[130]	T2DM	Forty individuals with T2DM and 60 HC.	3-indoxyl sulfate, glycerophospholipids, free fatty acids, and bile acids	Blood	NMR, MS
[131]	T2DM	KORA S4: 91 dT2D and 1206 non-T2DM (866 NGT, 102 i-IFG, 238 with IGT). KORA F4: 876 non-diabetic (91 developed T2DM). of 641 individuals with NGT at baseline (118 developed IGT))	Three significant metabolites (glycine, lysophosphatidylcholine (LPC) (18:2), and acetylcarnitine) had changed levels in IGT individuals as compared to those with normal glucose tolerance.	Serum	LC and flow injection analysis–MS
[132]	T2DM	T2DM (n = 800) and HC (n = 2282).	This study identified sugar metabolites, amino acids, and choline-containing phospholipids to be independently associated with risk of T2D.	Serum	Flow injection analysis tandem mass spectrometry

Ref	Disease / Treatment	Training Size	Biomarkers	Biological Matrix	Analytical Platform
					(FIA) MS/MS
[133]	T2DM	T2DM (n = 115), IFG (n = 192) and HC (n = 1897).	Forty-two metabolites from three major fuel sources, carbohydrates, lipids and proteins are robust risk factors for the development of both IFG and T2D. The branched-chain keto-acid metabolite 3-methyl-2-oxovalerate was the strongest predictive biomarker for IFG after glucose	Plasma and Urine	UHPLC-LTQ/MS and GC-MS
[134]	T2DM	24 IFG, 27 T2DM, and 60 ND	fructose, α-hydroxybutyrate, alanine, proline, phenylalanine, glutamine, BCAA (leucine, isoleucine,andvaline), low carbonnumberlipids (myristic, palmitic,andstearic acid), pyroglutamic acid, glycerophospohlipids, and sphingomyelins.	Serum	MS
[135]	IR	IR (n = 30) and non-IR (n = 30)	47 metabolites were found to be significantly different. Bile acids exhibit the greatest changes.	Serum	LC-MS, GC-MS, CE–MS.
[136]	T2DM coronary heart disease (T2DM-CHD)	15 HC, 13 CHD, 15 T2DM and 28 T2DM-CHD.	About 11 and 12 representative metabolites of CHD and T2DM were identified, respectively, mainly including alanine, arginine, proline, glutamine, creatinine and acetate.	Plasma	NMR
[137]	T2DM		BCAAs, AAAs, and acylcarnitines are strongly associated with early IR.	Blood, Saliva	MS, NMR
[128]	T2DM and Prediabetes	8,000 individuals (1,940 had T2DM)	Several blood amino acids appear to be consistently associated with the risk of developing T2DM.	Blood (plasma, serum) or urine	MS, NMR, HILIC.
[138]	T2DM	80 older people with late-onset T2DM and 79 older controls without T2DM.	Lower levels of phenylalanine, acetylhistidine, and cAMP were found in urine samples of late-onset T2DM subjects. Elevated levels of 5′-methylthioadenosine (MTA) was found in the urine of older people with T2DM.	Urine	UPLC-MS
[139]	T2DM	2776 participants (controls = 2564, cases = 212)	lipoprotein sub-fractions, certain triglycerides, amino acids	Blood	NMR, MS
[140]	T1DM, T2DM	DM (n = 26, type 1 (n = 8) and T2DM (n = 18))	Six important biomarkers were ethanol, acetone, methoxyacetone, hydroxyurea, isobutyraldehyde,	Earwax	GC-MS

Ref	Disease / Treatment	Training Size	Biomarkers	Biological Matrix	Analytical Platform
		and HC (n = 33).	and acetic acid. Methoxyacetone was the only biomarker able solely to perfectly discriminate between diabetes types 1 and 2.		
[141]	T2DM	503 cases, 503 Controls.	PCs containing odd-chain fatty acids (C19:1 and C17:0) and 2-hydroxyethanesulfonate were associated with the likelihood of developing T2DM.	Plasma	LC-MS
[142]	T2DM	160 incident and 144 prevalent cases with T2DM and 304 controls.	Several LPIs and NEFAs were associated with the risk of T2DM.	Serum	LC-MS/MS, GC MS/MS
[143]	T2DM	1939 participants with available metabolic profiles and without DM, cardiovascular disease, or cancer at baseline.	Serum levels of alanine, arginine, isoleucine, proline, tyrosine, valine, hexose and five phosphatidylcholine diacyls were positively associated with T2DM risk. In contrast, lyso-phosphatidylcholine acyl C17:0 and C18:2 and other glycerophospholipids were negatively associated with T2DM risk.	Serum	AbsoluteIDQ TM p180 kit, LC–MS/MS
[144]	T2DM	2939 participants with metabolomics data and without prevalent diabetes (1126 T2DM)	Food additive (erythritol) and compounds involved in amino acid metabolism [isoleucine, leucine, valine, asparagine, 3-(4-hydoxyphenyl)lactate] and glucose metabolism (trehalose)	Serum	Waters ACQUITY UPLC, ThermoFisher Scientific Q-Exactive MS
[145]	T2DM	11,896 individuals (392 T2DM)	Branched-chain and aromatic amino acids, triacylglycerol within VLDL particles, linoleic n-6 fatty acid, and non-esterified cholesterol.	Serum	NMR
[146]	T2DM and IR	30 adults of normal weight, 26 obese adults, and 16 adults newly diagnosed with T2DM	IR potential biomarkers: amino acids (Asn, Gln, and His), methionine (Met) sulfoxide, 2-methyl-3-hydroxy-5-formylpyridine-4-carboxylate, serotonin, L-2-amino-3-oxobutanoic acid, and 4,6-dihydroxyquinoline. T2DM was associated with dysregulation of 42 metabolites, including amino acids, amino acids metabolites, and dipeptides.	Serum	(CIL) LC-MS

Ref	Disease / Treatment	Training Size	Biomarkers	Biological Matrix	Analytical Platform
[147]	T2DM	789 participants of the PIVUS study (108 prevalent cases of T2DM) and 635 participants of the ULSAM study (89 cases of prevalent T2DM).	3-hydroxyundecanoyl-carnitine	Urine	UPLC-MS
[148]	T2DM		Direct association of BCAA and an inverse association of glycine with T2DM.		
[5]	T2DM	The number of participants ranged from 100(20) to 27296(21)	Isoleucine, leucine, valine, tyrosine, phenylalanine, glutamate, alanine, valerylcarnitine (C5), palmitoylcarnitine (C16), palmitic acid, and linoleic acid were associated with higher T2DM risk. However, serine, glutamine, and lysophosphatidylcholine C18:2 decreased the risk of T2DM.	plasma, serum, urine	MS, NMR
[149]	T2DM	100 Patients (obese non-T2DM (n = 50) and obese T2DM (n = 50)) UAE nationals aged between 18 - 60 year.	BCAAs, trimethylamine N-oxide, β-hydroxybutyrate, trimethyl uric acid, and alanine. lysophosphatidylcholines, phosphatidylcholines, acylcarnitine, amino acids and sphingomyelins;Lyso.PC.a.C18.0,PC.ae.C 34.2, C3.DC..C4.OH, glutamine and SM.C16.1, are the most significant metabolites.	Blood	NMR, FIA-MS/MS, and LC-MS/MS.

3.4 Metabolomics and Demographics Variables

Several factors have been revealed to have an association with the incidence of diabetes, bolstered by various epidemiologic and experimental research. Risk factors can be classified into two categories: non-modifiable, such as age, gender, ethnicity, family history, or modifiable such as BMI, diet, exercise, risk factors.

Different metabolomics studies show the significant impact of sex and age on metabolite profiles [158-166]. Their cross-sectional designs limit these studies, yet they are informative. It is vital to apply a longitudinal research design that can depict age-

associated phenomena to evaluate aging's metabolomics, mainly because of the high variability of metabolites [167].

It is stated that age is recognized to be the single major risk factor of the most widespread diseases in developed countries [168]. Building a knowledge of the metabolome variation with age could further unveil the mechanisms by which age impacts disease risk. It could also enable discovering high-risk metabolomic profiles suggestive of specific diseases' early stages [169]. Therefore, longitudinal plasma samples from the Wisconsin Registry for Alzheimer's Prevention were used to investigate the function of age in metabolomics [169]. The results show that 1,097 metabolites were tested, 623 (56.8%) were associated with age, and 695 (63.4%) with sex after correcting for multiple testing. The levels of most metabolites are significantly affected by age and sex, and sex differentially influences the levels and trajectories of many metabolites. The study's significance underlines the importance of integrating age and sex in the design and analysis of metabolomics studies and proposes a richer insight into the aging process that could tell many novel hypotheses regarding the role of metabolites in healthy and accelerated aging.

The Baltimore Longitudinal Study of Aging was employed to identify plasma metabolites predictive of change in gait speed over time [170]. Gait speed measures lower extremity physical performance in older adults and predicts disability and mortality. BLSA is a follow-up study of 50.5 months in 504 adults aged 50 years and above. Results show that of 148 plasma metabolites (amino acids, biogenic amines, hexoses, glycerophospholipids) measured, eight were significantly associated with gait speed at baseline, independent of age and sex. It is concluded that Low plasma LPC 18:2, which has previously been shown to predict IGT, IR, T2DM, coronary artery disease, and memory impairment, is an independent predictor of decline in gait speed in older adults.

It is worth mentioning that [157] summarizes major classes of metabolites affected by age, sex, and BMI. Therefore, it is critical to study the impact of different demographic features on metabolites in the UAE population.

3.5 Metabolomics in Diabetic Kidney Disease

Recently, the investigation of DKD via metabolomics has been of primary interest [171, 172]. However, despite the increased interest in metabolomics in DKD patients [173,

174], more studies need to be conducted in such a manner. Specifically, studies on people with diabetes under hemodialysis have been rare.

Several studies have shown potential biomarkers of DKD. The primary metabolites were products of lipid metabolism (such as esterified and nonesterified fatty acids, carnitines, phospholipids), branch-chain amino acid and aromatic amino acid metabolism, carnitine, and tryptophan metabolism, nucleotide metabolism (purine, pyrimidine), and the tricarboxylic acid cycle or uraemic solutes [175-178]. Moreover, mitochondrial function and fatty acid oxidation are crucial in the DKD progress [172, 179]. However, these studies demonstrated substantial variations in the metabolomic profiles, perhaps due to differences in geography, ethnicity, sample selection, and analytical platform.

The metabolomic profile of DKD under hemodialysis from the middle eastern populations is unknown. Therefore, the second study explores the metabolomic profile of diabetic and non-diabetic UAE citizens undergoing hemodialysis to uncover the potential novel biomarkers in this population. However, diabetic medication intake for dialysis patients affects their metabolic profiling. Therefore, we also analyzed the data based on the available HbA1c values.

Metabolomics is promising in the pharmaceutical field and clinical research. However, due to the complexity and high throughput data generated from such experiments, data mining and analysis are significant challenges for researchers in the field. Thus, several efforts were achieved to develop a complete workflow that helps researchers analyze data. The following sections review the state-of-the-art computer-aided tools and databases in metabolomics established in recent years.

3.6 Metabolomics Databases

The ever-growing quantity of experimental and computational chemical data requires consideration of storing, accessing, and manipulating this vast amount of information. Today, hundreds of database projects are created and annotating biological knowledge; each is dedicated context, as shown in Figure 3-3.

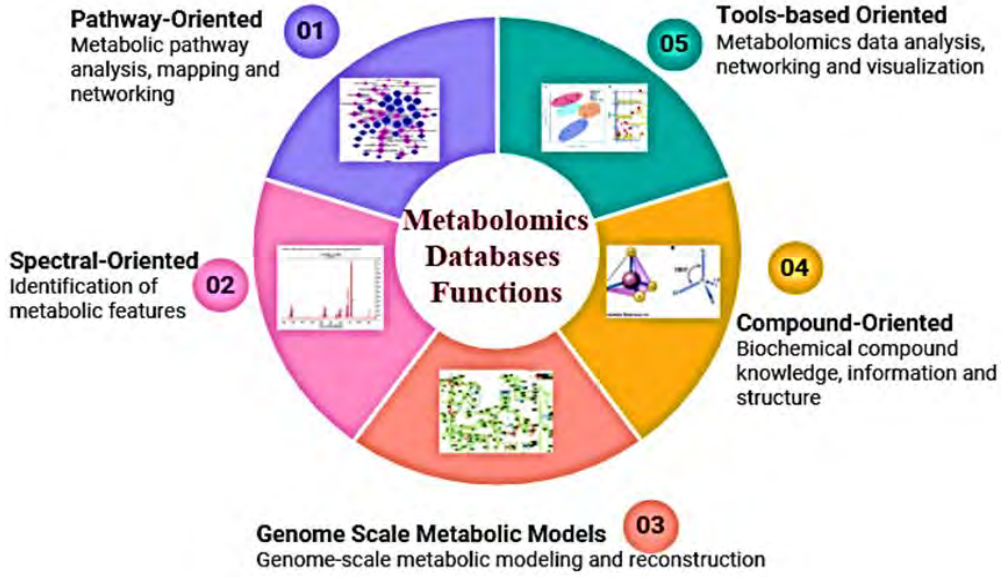


Figure 3-3: Metabolomics databases multifunctional tasks.

As a result, the database's current catalog is robust and diverse, including organism focus, curation approach, type of pathways, interactions covered, and other differences. In addition, many databases are available to researchers for data mining and sharing consistent chemical data for various purposes. For example, all pathway search tools depend on a database from which biochemical reactions and molecules can be enlisted to comprise the pathway of interest. This section discusses the databases related to various metabolite annotation, metabolism, and metabolomics workflows.

The Reactome Knowledgebase [180] is a curated database of pathways and reactions in human biology, cross-referenced with several resources, as essential literature and different pathway-related databases. It aims its manual annotation effort on a single species, Homo sapiens, and applying a single consistent data model across all domains of biology. The Reactome defines a reaction as any event in biology that changes the state of a biological molecule. Binding, activation, translocation, degradation, and classical biochemical events involving a catalyst are all reactions. It provides molecular details of signal transduction, transport, DNA replication, metabolism, and other cellular processes. It contains 2,546 human pathways and 1,940 small molecules [180].

BioCyc [181] is an encyclopedic reference to a collection of 19494 Pathway and Genome Databases for model eukaryotes and thousands of microbes and software tools for exploring them. In addition, BioCyc comprises curated data from 130,000 publications. The MetaCyc and EcoCyc databases are freely available via BioCyc.

However, access to the remaining BioCyc databases needs a paid subscription such as HumanCyc (HumanCyc.org) [182].

MetaCyc [183] is a broad reference database of metabolic pathways and enzymes from all fields of life. It includes 2937 pathways obtained from 3295 different organisms., making it the most extensive curated collection of metabolic pathways [183].

EcoCyc [184] is a scientific database for Escherichia coli K-12 MG1655. The EcoCyc project performs literature-based curation of its genome, transcriptional regulation, transporters, and metabolic pathways. New and improved data analysis and visualization tools include an interactive metabolic network explorer, a circular genome viewer, and many upgrades to the speed and usability of existing tools [184]. It mainly focuses on metabolic pathways and signaling.

Metabolite Network of Depression Database (MENDA) [185] is a broad metabolite-disease association database that integrates all existing knowledge and datasets of metabolic characterization in depression. In addition, study and tissue type, organism, category of depression, sample size, platform (MS-based, MRS, NMR), and differential metabolites are provided.

BiGG Models [186] is Biochemical, Genetic, and Genomic knowledge base of genome-scale metabolic network reconstructions. BiGG Models include more than 75 high-quality, manually curated genome-scale metabolic models. It also delivers a broad application programming user interface for accessing BiGG Models with modeling and analysis tools. In addition, reaction identifiers, metabolite identifiers, and pathway visualization were formalized in BiGG Models.

KEGG [46] is one of the most complete and widely used databases. It is a manually curated resource integrating eighteen databases categorized into systems, genomic, chemical and health information.

The BRaunschweig ENzyme Database (BRENDA) enzyme database [187] contains comprehensive functional enzyme and metabolism data such as measured kinetic parameters. The main part has more than 5 million data for almost 90000 enzymes. In addition, BRENDA offers easy access to enzyme information from quick to advanced searches, text- and structured-based queries for enzyme-ligand interactions, word maps, and visualization of enzyme data.

PubChem [47] is the world's most extensive collection of freely accessible chemical information from more than 750 data sources. It stores information in three primary categories: compounds, substances, and bioactivities. In addition, several research areas use PubChem as a big data source, including machine learning and data science studies for virtual screening, drug repurposing, chemical toxicity prediction, drug side effect prediction, and metabolite identification. Furthermore, PubChem provides chemical and physical properties, biological activities, safety and toxicity information, patents, literature citations, and more.

ChEBI [51] is a freely accessible dictionary of molecular entities focused on small chemical compounds. The HMDB [45] is comprehensive reference information about human metabolites and their related biological, physiological, and chemical properties. To date, HMDB has a 220945 total number of metabolites. ChemSpider [188] is a free chemical structure database delivering quick text and structure search access to more than 100 million structures from hundreds of data sources.

MetaboLights [189] is a database for metabolomics studies, raw experimental data, and associated metadata. MetaboLights is cross-species, cross-technique, and covers metabolite structures and their reference spectra and their biological roles, locations and concentrations, and experimental data from metabolic experiments. Users can upload their research datasets into the MetaboLights Repository. These researches are then automatically given a stable and unique identifier that can be used for publication reference.

The Metabolomics Workbench [190] is a public repository for metabolomics metadata and experimental data spanning various species and experimental platforms, metabolite standards, metabolite structures, protocols, tutorials, training material, and other educational resources. It can integrate, analyze, track, deposit, and disseminate big heterogeneous data from many MS- and NMR-based metabolomics studies. It also covers more than 20 different species, including humans and other mammals, plants, insects, invertebrates, and microorganisms.

SMPDB [53] is a comprehensive, interactive, visual database containing over 48000 discovered pathways. Most of these pathways don't exist in any other pathway database. SMPDB help in pathway interpretation and pathway discovery in metabolomics, transcriptomics, proteomics, and systems biology.

MetSigDis [191] is a free web-based tool that offers a comprehensive metabolite alterations resource in various diseases. The database deposited 6849 curated relationships between 2420 metabolites and 129 diseases across eight species involving Homo sapiens and model organisms.

Virtual Metabolic Human [192] is a web-based database capturing current knowledge of human metabolism within five interlinked resources including, Human metabolism, Gut microbiome, Disease, Nutrition, and ReconMaps. The VMH's unique features are (i) the hosting of the metabolic reconstructions of human and gut microbes amenable for metabolic modeling; (ii) seven human metabolic maps for data visualization; (iii) a nutrition designer; (iv) a user-friendly webpage and application-programming interface to access its content; (v) user feedback option for community engagement and (vi) the connection of its entities to 57 other web resources.

WikiPathways [193] is a reliable and rich pathway database that captures biological pathways' collective knowledge. By providing a database in a curated, machine-readable way, omics data analysis and visualization is enabled.

The relational database of Metabolomics Pathways (RaMP) [194] is a public database to combine biological pathways from the KEGG, Reactome, WikiPathways, and the HMDB. RaMP maps genes and metabolites to biochemical/disease pathways and can readily be integrated into other existing software. It can be used as a stand-alone resource or incorporated into other tools.

Pathway Commons [195] is one of the most extensive composite databases. It is an integrated resource of publicly available information about biological pathways, including biochemical reactions, assembly of biomolecular complexes, transport and catalysis events, and physical interactions involving proteins, DNA, RNA, and small molecules (e.g., metabolites and drug compounds). A list of commonly used metabolomics databases and their main features can be found in Table 3-2.

A variety of databases stands as a metabolomics datasets repository. To mention some, BioMagResBank (BMRB) [52] is a public repository for NMR spectroscopy data from proteins, peptides, nucleic acids, and other biomolecules. In addition, Golm Metabolome Database (GMD) [196] provides data sets for biologically quantified active metabolites and text search capabilities for GC-MS data. Moreover, the Mass Spectral Library [197] extensively collects EI MS, MS/MS, Replicate spectra, and

Retention index data sets. Finally, the Spectral Database System (SDBS) [198] is a spectral database for organic compounds and has various MS, NMR, IR, Raman, ESR data sets.

Taken all together, Pathguide [199] is a necessary initial step for considering the prospect of pathway databases. Pathguide is a meta-database that contains information about 702 biological pathway-related databases and molecular interaction-related databases. For example, the Pathguide categories include metabolic pathways, signaling pathways, pathway diagrams, transcription factor targets, gene regulatory networks, genetic interactions networks, protein–compound interactions, protein sequence-focused, protein-protein interactions, etc.

Despite the emerging number of chemical databases, the significant challenge for this expansion is the incompetence to utilize metabolite and reaction information from different databases such as KEGG, BRENDA, MetaCyc because of representation inconsistencies and, duplications and errors. In addition, the same metabolite is found with multiple names across databases and models, which slows down collating information from various data sources. Therefore, researchers designed the MetRxn database [200], Rhea [201], and RefMet [202] to standardize reaction and metabolite names. Additions and modifications to databases are made regularly to increase the quality and coverage of their biological knowledge. Some databases can update their information frequently to sustain pace with discoveries. For instance, the KEGG database [46] revises its data weekly; however other databases do it less often. The preference of databases should consider the relative sizes, degree of overlap, and scope. For instance, KEGG contains significantly more compounds than MetaCyc, whereas MetaCyc contains more reactions and pathways than KEGG. For example, pathway sets can differ between databases in many ways, including the number of pathways present, the size of pathways, how pathways are curated (manually or computationally, or a combination of both), pathway boundaries, and the organisms supported [203]. However, the interpretation of metabolomics data has been challenging as understanding the connections between dozens of altered metabolites has often relied on researchers' biochemical knowledge and speculations. However, modern biochemical databases provide information about metabolism's interrelations, automatically polling using metabolomics secondary analysis tools, i.e., mathematical, and computational tools. Table 3-2 shows variety of available databases online. The

second column shows the kind of species pursued by each database. Additionally, in the sixth column, the links to websites are supplied.

Table 3-3: Summary of metabolomics databases.
YOR, Year of release.

Database	Organisms	Database descriptions	Coverage	Accessibility	Link	Y.O.R	Ref.
Reactome Knowledgebase	Homo sapiens	It contains visualization, interpretation, and analysis of pathway knowledge. Available tools: SkyPainter, PathFinder, BioMart, Reactome Gene Set Analysis (ReactomeGSA) and Reactome IDG Portal.	**Human Pathways**:2546 **Reactions**:13890 **Proteins**:1020 **Small Molecules**:1940 **Drugs**:507	Free	Reactome.org	2005	[180]
BioCyc	Eukaryotes Bacteria and Archaea.	It is a comprehensive reference containing listed data from 130000 publications—available tools: Pathologic, Genome browser, Pathway Tools, BLAST search and SmartTables.	**Pathway/Genome Databases (PGDBs):**19494 **Archaea**: 465 databases **Bacteria**: 18956 databases **Eukaryota**: 37 databases **MetaCyc**: Metabolic Encyclopedia	EcoCyc and MetaCyc databases: free access. Others: Paid subscription	Biocyc.org	1997	[181]
MetaCyc	Eukaryotes Bacteria and Archaea.	It serves as a comprehensive reference to metabolic pathways and enzymes. Available tools: Pathologic, Genome browser, BLAST search, Pathways Tools, Google™.	**Multi-organisms**: 3295 **Metabolic pathways**:2937 **Enzymatic reactions**:17310	Free	MetaCyc.org	1999	[183]
EcoCyc	Bacterial organism: Escherichia coli K-12 MG1655	It contains Metabolic Network Explorer, Circular Genome Viewer	**Genes**:4518 **Enzymes**:1682 **Metabolic reactions**:2151	Free	EcoCyc.org	1995	[184]
BIGG Models	Eukaryotes, Prokaryotes, and Photosynthetic Eukaryotes.	It provides Pathway visualization with Escher. It also offers Standardized identifiers for metabolites, reactions, and genes.	It contains more than 75 high-quality manually-curated genome-scale metabolic models.	Free	BIGG.ucsd.edu	2007	[186]

Database	Organisms	Database descriptions	Coverage	Accessibility	Link	Y.O.R	Ref.
KEGG	Eukaryotes Bacteria and Archaea.	PATHWAY database, KEGG NETWORK database, KO annotation and taxonomy, Drug information, and Virus-cell interaction. Available tools: KEGG Atlas, KegHier, KegArray, KegDraw, KegTools, KEGG2, KEGG API.	**KEGG organisms:** 7760 (**Eukaryotes:** 695, **Bacteria:**6694**, Archaea:**371). **KEGG modules:** 456 **Reaction modules:**46	Free	www.kegg.jp/	1995	[46]
BRENDA	Eukaryotes Bacteria and Archaea.	It comprises disease-related data, protein sequences, 3D structures, genome annotations, ligand information, taxonomic, bibliographic, and kinetic data.	**Number of different enzymes:** 8197	Free	www.brenda-enzymes.org	1987	[187]
PubChem	Eukaryotes Bacteria and Archaea	It provides chemical and physical properties, biological activities, safety and toxicity information, patents, literature citations, and more. Available tools: PubChem Structure Editor, Entrez, PubChem3D, PubChem Download Facility, ToxNet .	**Compounds**: 110 million, **Substances**:277 million, **Bioactivities**: 293 million.	Free	PubChem.ncbi.nlm.nih.gov	2004	[47]
ChEBI	Eukaryotes Bacteria and Archaea	It is a database and ontology containing information about chemical entities of biological interest.	**Annotated compounds**: 59708	Free	www.ebi.ac.uk/chebi	2010	[51]
HMDB	Homo sapiens	It is a human metabolomics database. It has spectral and pathway visualization tools. Available tools: Data Extractor, ChemSketch, BLAST search, MetaboCard, MS and NMR spectral search utility, MetaboLIMS.	**Annotated metabolite entries**: 217920	Free	https://hmdb.ca	2007	[45]

Database	Organisms	Database descriptions	Coverage	Accessibility	Link	Y.O.R	Ref.
ChemSpider	Eukaryotes Bacteria and Archaea	It is a chemical structure database.	**Chemical entities**:114 Million	Free	chemspider.com	2007	[188]
MetaboLights	Eukaryotes Bacteria and Archaea	It is an open-access database repository for cross-platform and cross-species metabolomics research.	**Different organisms:** 6510 **Reference compounds:**27475 **Metabolite annotation features:**2016457	Free	https://www.ebi.ac.uk/metabolights	2012	[189]
Metabolomics Workbench	Eukaryotes Bacteria and Archaea	It is a repository for metabolomics data and metadata and provides analysis tools and access to metabolite standards, protocols, tutorials, training, and more.	**Discrete structures:**136000 **Genes:**7300 **Proteins:**15500	Free	metabolomicsworkbench.org	2016	[190]
SMPDB	Eukaryotes Bacteria and Archaea	It is a pathway database for different model organisms such as humans, mice, E. coli, yeast, and Arabidopsis thaliana.	**Pathways Number:** 48690 **Metabolites Number (non-redundant):** 55700	Free	https://smpdb.ca/	2009	[53]
MetSigDis	Homo sapiens, Rat, Mouse, Drosophila melanogaster, Triatomine, Mice, Pig, and Mus musculus.	It is a manually curated resource that aims to provide a comprehensive resource of metabolite alterations in various disease.	**Curated relationships:6,849 Metabolites: 2,420 Diseases:129 Species: 8**	Free	http://www.bio-annotation.cn/MetSigDis/	2017	[191]
Virtual Metabolic Human	Homo sapiens	It captures human and gut microbial metabolism information and links it to hundreds of diseases and nutritional data.	**Reactions:19313 Metabolites: 5607 Human genes:3695 Diseases:255 Foodstuff:8790**	Free	www.vmh.life	2018	[192]
Pathway Commons	Eukaryotes Bacteria and Archaea	It aims to collect and disseminate biological pathway and interaction data	**Pathways:5772 Interactions: 2424055 Databases:22**	Free	https://www.pathwaycommons.org		[195]

Database	Organisms	Database descriptions	Coverage	Accessibility	Link	Y.O.R	Ref.
WikiPathways	Eukaryotes Bacteria and Archaea	It is a public, collaborative platform devoted to the curation of biological pathways	**Human genes: 11532 Number of pathways: 3013**	Free	wikipathways.org	2008	[193]
RaMP	Eukaryotes Bacteria and Archaea	It is a multi-database integration approach for gene/metabolite enrichment analysis providing interactive tables of query results, interactive tables of PA results, and clustering of enriched pathways by pathway similarity	**Pathways: 51,526 (from KEGG, Reactome, SMPDB, and WikiPathways) Genes: 23,077 Metabolites: 113,725**	Free	https://github.com/mathelab/RaMP-DB/ or https://github.com/mathelab/RaMP-DB/inst/extdata/		[194]
MENDA	Organisms include: Human, Rat, Mouse, and Non-human primates.	It is a comprehensive metabolic characterization database for depression.	Differential expressed metabolites: 5675. **(Humans**:1347 **Rat**:3127 **Mouse**:1105 **Non-human primates**:96)	Free	Menda.cqmu.edu.cn:8080/index.php	2020	[185]

3.7 Metabolomics Computer-Aided Tools

Python (https://www.python.org/), R [204], and other programming languages empower and facilitate various tools to implement integrated workflows. Independent computational methods for conducting statistics, enrichment, visualization, and contextualization should be combined into integrated workflows [205]. These workflows should be customized and compatible with the study designs to attain complete and significant information from the metabolomics datasets. Mathematical methods are helpful for molecular biomarker detection. However, statistical tests, such as t-test, significance analysis of microarrays (SAM), and eBayes, are commonly used to extract dysfunctional molecules from large-scale expression data, integrated as an essential analytical step in many biomarker identification pipelines. In addition, several novel computational tools have been established as secondary analysis tools to enable metabolomics researchers to grasp the powers of their data and produce farther-reaching biological conclusions than ever before. This section explains the functionality and use of various analysis tools.

The MarVis-Suite [206] (Marker Visualization) toolbox for interactive ranking, filtering, combination, clustering, visualization, and functional analysis of data sets containing intensity-based profile vectors, as found, e.g., from MS, microarray, or RNA-seq experiments.

MetExplore [207] offers an easy-to-use complete online solution comprised of interactive tools for metabolic network curation, network exploration, and omics data analysis. MetExplore holds the concepts of metabolic networks and significantly improves multi-omics data analysis.

Pathway Activity Profiling [208] (PAPi) compares metabolic pathway activities from metabolite profiles. PAPi can reach the activity of metabolic pathways under different conditions, which provides excellent support for hypothesis generation and facilitates biological interpretation.

Metabolites Biological Role (MBROLE) [209] is a server that performs functional enrichment analysis of a list of chemical compounds derived from a metabolomics experiment, which allows this list to be interpreted in biological terms. MBROLE analyzes a wide variety of functional annotations that describe many different aspects of the chemistry and biology of chemical compounds; these include pathways and sub-pathways, interactions with enzymes, proteins and other types of molecules, physiological locations, chemical classifications and taxonomies, and biological roles, uses, and applications. MeltDB 2.0 [210] is a next-generation web application addressing storage, sharing, standardization, integration, and analysis of metabolomics experiments.

MetaboAnalyst version 5.0 [211] is a fully automated web interface to bridge raw data to functional insights for global metabolomics based on high-resolution mass spectrometry (HRMS). MetaboAnalyst performs optimized peak detection, alignment, and annotation tasks for LC-MS data generated in global metabolomics. The key features of MetaboAnalyst are that it includes: (1) MetaboAnalystR package in R environment, (2) large libraries for metabolite sets and metabolic pathways, (3) metabolomic biomarker metanalysis, (4) integration of multi-omics data through knowledge-based network analysis and visualization, and (5) easy and free accessible tool.

Metabolite pathway enrichment analysis (MPEA) [212] is a metabolomics pathway enrichment tool for visualization and biological interpretation. However, MPEA is limited to top-down/bottom-up analysis.

MetaP-server [213] is an easy-to-use web-server-based for metabolomics data analysis. It covers data acquisition to biological interpretation: (i) data quality checks, (ii) estimation of reproducibility and batch effects, (iii) hypothesis tests for multiple categorical phenotypes, (iv) correlation tests for metric phenotypes, (v) optionally including all possible pairs of metabolite concentration ratios, (vi) PCA, and (vii) mapping of metabolites onto colored KEGG pathway maps.

Mass TRanslator into Pathways (MassTRIX) [214] annotates metabolites in high precision mass spectrometry data. It marks the identified chemical compounds on KEGG pathway maps using the KEGG/API. In addition, selected genes or enzymes can be highlighted, e.g., to represent information on gene transcription or differences in the gene complement of different bacterial strains.

Pathos [215] is a web-based tool to analyze raw or processed metabolomics mass spectra and demonstrate the metabolites identified and alterations in their experimental abundance within the context of their associated metabolic pathways. Pathos is limited to specific organism databases.

PaintOmics 3 [216] is a web-based tool for the integrated visualization of multiple Omic data types onto KEGG pathway diagrams. PaintOmics 3 combines server-end capabilities for data analysis with the potential of modern web resources for data visualization, delivering researchers with a robust framework for interactive exploration of their multi-omics information.

IMPaLA [217] is a web-based tool for the joint PA with expression (genes/proteins) and metabolite data. It performs over-representation or enrichment analysis with user-specified lists of metabolites and genes using over 3000 pre-annotated pathways from 11 databases.

MetaMapR [218] is an open-source, web-based, or desktop software implemented in the R programming language. It integrates enzymatic transformations with metabolite structural similarity, mass spectral similarity, and empirical associations to generate well-connected metabolic networks.

The Layered Enrichment Analysis of Pathways (LeapR) [219] is a framework to measure biological pathway activity using various statistical tests and data sources, allowing facile integration of multisource data.

PAthway NEtwork Visualizer (PANEV) [220] is an R package set for gene/pathway-based network visualization. Utilizing KEGG, it visualizes genes within a network of multiple levels of interconnected upstream and downstream pathways. The network graph visualization helps to interpret functional profiles of a cluster of genes. However, PANEV is a KEGG-based tool that can be considered a limitation because of KEGG's lack of or incomplete information.

PathfindR [221] is an R package using protein-protein interaction information and for active-subnetwork-oriented pathway enrichment analyses for class comparison omics experiments. It also provides functionality for clustering the resulting pathways.

Ingenuity Pathway Analysis [222] is A comprehensive visualization software/database search tool for finding functions and pathways for specific biological states. IPA helps understand complex omics data and perform insightful data analysis and interpretation by placing experimental results within the context of biological systems. Its pathway focuses on protein-protein interactions, protein-compound interactions, metabolic, signaling, gene regulation, and diagrams.

iPath3.0 [223] is a free web-based tool for visualization, customization, and analysis of various KEGG cellular pathways. Version 3 could deal with metabolic pathway, regulatory pathway, and biosynthesis of secondary metabolites.

ReactomePA [224] is a free R/Bioconductor package providing enrichment analyses, including hypergeometric tests and gene set enrichment analyses. A functional analysis can be applied to the genomic coordination obtained from a sequencing experiment to analyze genomic loci's functional significance, including cis-regulatory elements and non-coding regions. In addition, ReactomePA provides several visualization functions to produce highly customizable, publication-quality figures.

MetExploreViz [225] is an open-source web component for visualizing metabolic networks and pathways and offers a flexible solution to analyze omics data in a biochemical context.

Recon3D [226] is a computational resource that includes three-dimensional (3D) metabolite and protein structure data and enables integrated analyses of metabolic functions in humans. Recon3D represents the most comprehensive human metabolic network model to date, accounting for 3,288 open reading frames (representing 17% of functionally annotated human genes), 13,543 metabolic reactions involving 4,140 unique metabolites, and 12,890 protein structures. These data provide a unique resource for investigating molecular mechanisms of human metabolism.

ChemRICH [227] is a statistical enrichment approach based on chemical similarity rather than sparse biochemical knowledge annotations. ChemRICH utilizes structure similarity and chemical ontologies to map all known metabolites and name metabolic modules. Unlike pathway mapping, this strategy yields study-specific, non-overlapping sets of all identified metabolites.

KEGGREST[228] is an R package employed to build an adjacency matrix that linked the dataset's metabolites with their corresponding KEGG pathways. First, one is assigned if the metabolite is part of that particular pathway, or 0 if not. Then five metabolites of each pathway were randomly sampled.

MetaX [229] offers several functions: peak picking and annotation, data quality assessment, missing value imputation, data normalization, univariate and multivariate statistics, power analysis and sample size estimation, receiver operating characteristic analysis, biomarker selection, and pathway annotation, correlation network analysis, and metabolite identification. It is available as a web-based interface and R package (http://metax.genomics.cn).

Biomarker Discovery by Machine Learning (BioDiscML) [230] is a biomarker discovery tool that exploits various feature selection procedures to produce signatures associated with machine learning models that efficiently predict a specified outcome. BioDiscML employs a large selection of machine learning algorithms to choose the best combination of biomarkers for predicting categorical or continuous outcomes from highly unbalanced datasets. BioDiscML can implement data pre-processing, feature selection, model selection, and performance evaluation. The software tool is developed in JAVA 8 language and uses the Weka 3.8 machine learning library. It outperforms recent tools for discovering biomarkers' signatures.

ASICS [231] is an R package that contains a complete workflow to analyze spectra from NMR experiments. It includes an automatic approach to identifying and quantifying metabolites in a complex mixture spectrum and uses the quantification results in untargeted and targeted statistical analyses. ASICS has algorithm limitations: the difficulty in detecting the metabolites with low concentrations or their peaks, all located in a region with a high density of peaks.

3Omics [232] is a web-based systems biology visualization tool integrating human transcriptomic, proteomic, and metabolomic data. It generates inter-Omics correlation networks to visualize relationships in data for time or experimental conditions for all transcripts, proteins, and metabolites.

To take a glimpse at such tools, a study [239] examined about 100 metabolomics software resources, tools, databases, and other utilities that emerged or were enhanced in 2019. Similarly, around 85 metabolomics software resources, packages, tools, databases, and other utilities that appeared in 2020 are released in a recent study [233]. Finally, Table 3-3 surveyed commonly used metabolomics tools in the literature.

Each of the available tools has strengths and weaknesses, and it should not come to the use of one over the other. The use of at least one enrichment analysis and one visualization/mapping tool is optional. Due to the complexity of metabolomics data, it is also essential to cautiously regard the results from the secondary analysis. For example, enrichment analysis can produce significant pathway hits from only one or two metabolites in a pathway. As such, careful scrutinization and logical biological interpretation of the data must be undertaken. With this in mind, metabolomics researchers should integrate secondary analysis into their studies as these beneficial results can be obtained rapidly [234]. The field of secondary analysis is coming into its own, and its steady growth will help enhance the success of the metabolomics approach. These cutting-edge bioinformatics analysis tools that are completely incorporated with various functions and are accessible and manageable by users who lack prior knowledge in programming are vital in metabolomics research. They will persist in enabling discoveries and more significant insights for increasing metabolomics researchers.

The dissertation performed two studies: (1) an untargeted metabolomics profile study for the T2DM Emirati population vs. healthy and (2) untargeted metabolomic profiling of Emirati dialysis patients with diabetes versus non-diabetic dialysis.

Table 3-4: Summary of computer-aided metabolomics.

Tool Name	Description	Input	Implementation	Accesibility	Databases Used	Link	Ref
MarVis-Suite	Metabolic pathways analysis and visualization	MS, microarray, or RNA-seq experiments	Web-based	Free	KEGG and BioCyc	http://marvis.gobics.de	[206]
MetExplore	Metabolic network and OMICs data analysis	Any	Web-based	Free	BioCyc- related	https://metexplore.toulouse.inra.fr/metexplore2/.	[207]
PAPi	Compare activity of metabolic pathway between sample types.	Any	R package	Free	KEGG	http://www.4shared.com/file/s0uIYWIg/PAPi_10.html	[208]
MBROLE	Enrichment analysis of metabolites annotations.	Any	Web-based	Free	KEGG, HMDB, PubChem, ChEBI, SMILES, YMDB, ECMDB, BioCyc-related, Rhea, UniPathway, LMSD, CTD, MeSH, MATADOR, DrugBank.	http://csbg.cnb.csic.es/mbrole2.	[209]
MetaboAnalyst 5.0	Metabolomics analysis platform, tutorials, and report analysis.	LC, GC raw spectra, MS, NMR peak list, and spectral bins.	Web-based, R package	Free	KEGG, HMDB, PubChem, ChEBI, RefMet and LIPID MAPS.	https://www.metaboanalyst.ca	[211]
MPEA	Pathway enrichment analysis.	Pre-annotated compounds or GC-MS-	Web-based	Free	KEGG, SMPDB and GMD.	http://ekhidna.biocenter.helsinki.fi/poxo/mpea/	[212]

Tool Name	Description	Input	Implementation	Accesibility	Databases Used	Link	Ref
		based MSTs					
PaintOmics 3	Compound mapping	Any	Web-based	Free	KEGG	www.paintomics.org	[216]
IMPaLA	Enrichment analysis.	Any	Web-based	Free	Reactome, KEGG, Wikipathways, HMDB, CAS, ChEBI, PubChem, SMPDB, NetPath, BIOCART, BioCyc.	http://impala.molgen.mpg.de	[217]
MetaMapR	Metabolic network mapping.	LC and GC raw spectra, MS and NMR peak list, and spectral bins.	Web-based or desktop software.	Free	KEGG and PubChem	http://dgrapov.github.io/MetaMapR/	[218]
LeapR	Enrichment analysis.	Any	R package	Free		https://github.com/biodataganache/leapR	[219]
PANEV	Gene/pathway-based network visualization	Any	R package	Free	KEGG	https://github.com/vpalombo/PANEV	[220]
PathfindR	Enrichment analysis.	Any	R package	Free	KEGG, Biogrid, v, IntAct,	https://cran.r-project.org/package=pathfindR	[221]
Ingenuity Pathway Analysis	Metabolic network mapping.	Any	Web-based, software	Paid	GO, KEGG, BIND	IPA, http://www.ingenuity.com	[222]
iPath3.0	Metabolic network mapping.	Compound IDs	Web-based	Free	KEGG, Uniprot, STRING, protein IDs, COGs, eggNOGs , NCBI gene identifiers, ChEBI and PubChem.	http://pathways.embl.de	[223]

Tool Name	Description	Input	Implementation	Accesibility	Databases Used	Link	Ref
ReactomePA	Enrichment analysis.	Any	R-package	Free	REACTOME	http://www.bioconductor.org/packages/ReactomePA	[224]
MetExploreViz	Metabolic network mapping.	Any	Web-based	Free	KEGG	http://metexplore.toulouse.inra.fr/metexploreViz/doc/	[225]
Recon3D	Network reconstruction	Any	Web-based	Free	KEGG, PDB, CHEBI, PharmGKB, UniProt	http://vmh.life	[226]
ChemRICH			Web-based and R-package	Free	NCBI BioSystems, PubChem, KEGG, BioCyc, Reactome, GO, and Wikipathways	www.chemrich.fiehnlab.ucdavis.edu) and www.github.com/barupal/chemrich	[227]
KEGGREST	A package provides a client interface to the KEGG REST server.	Compound IDs	R package	Free	KEGG	https://bioconductor.org/packages/release/bioc/html/KEGGREST.html	[228]
MetaX	Flexible and comprehensive software for processing metabolomics data	Raw peak intensity data	Web-based and R-package	Free	HMDB, KEGG, MassBank, PubChem, LIPID MAPS, MetaCyc, and PlantCyc	http://metax.genomics.cn).	[229]
BioDiscML	Biomarker discovery software that supports classification and regression problems.	Any	Stand-alone program	Free		https://github.com/mickaellecl ercq/BioDiscML.	[230]

Tool Name	Description	Input	Implementation	Accesibility	Databases Used	Link	Ref
3Omics	Web tool visualization of multi-omics data (transcriptomics, proteomics, and metabolomics)	Any	Web-based	Free	iHOP, KEGG, HumanCyc, DAVID, Entrez Gene, OMIM and UniProt	http://3omics.cmdm.tw	[232]
MeltDB 2.0	Web-based tool for statistical analysis and sets for enrichment analysis.	Raw GC/LC-MS spectra, processed spectra, compound IDs, and abundances.	Web-based, login required	Free	KEGG, ChEBI, GMD and CAS.	https://meltdb.cebitec.uni-bielefeld.de	[210]
MassTRIX	Compound mapping	MS spectra	Web-based	Free	KEGG, HMDB and LipidMaps.	www.masstrix.org	[214]
MetaP-server	Global statistical analysis	Compound IDs and sample metadata.	Web-based	Free	KEGG, HMDB, LIPID MAPS, PubChem and CAS.	http://metabolomics.helmholtz-muenchen.de/metap2/)	[213]
Pathos	Compound mapping	MS-spectra (raw *m/z*) and compound IDs (KEGG or MetaCyc IDs)	Web-based	Free	KEGG	http://motif.gla.ac.uk/Pathos/)	[215]

Chapter 4. Methodology

4.1 Introduction

This chapter depicts the chronological methodological stages followed in the dissertation. First, it covers ethical approval for biological samples acquisition and then analytical steps conducted in the SIMR lab. Finally, it describes data statistical and pathway analysis. Figure 4-1 records sequential methodological steps in our work.

4.2 Participant Inclusion and Ethical Statement

After obtaining ethical approval from the University Hospital Sharjah Ethics Research Committee (REF number: UHS-HERC-012-10062019), the study was conducted with adherence to the committee's research guidelines and regulations. The first comprehensive study has 92 subjects: 50 were diagnosed with T2DM, and the other 42 subjects have no known T2DM status and are referred to as a non-T2DM group. The second study has 36 patients, including 11 dialysis diabetic patients and 25 non-diabetic dialysis patients. However, the sample size is restricted by the available resources, such as individuals' willingness to participate and the cost of sample analysis. In addition, the selected individuals are located in Sharjah, UAE. Also, we acknowledge the lack of confounding variables such as stress.

All volunteers were supplied with an information sheet explaining study objectives, design, and confidentiality, and written informed consent was obtained from all study participants. One hundred twenty-eight blood specimens, 4 ml each, were collected in sterile containers. The samples were stored immediately at 4°C for short-term storage or –80 ºC for long-term storage.

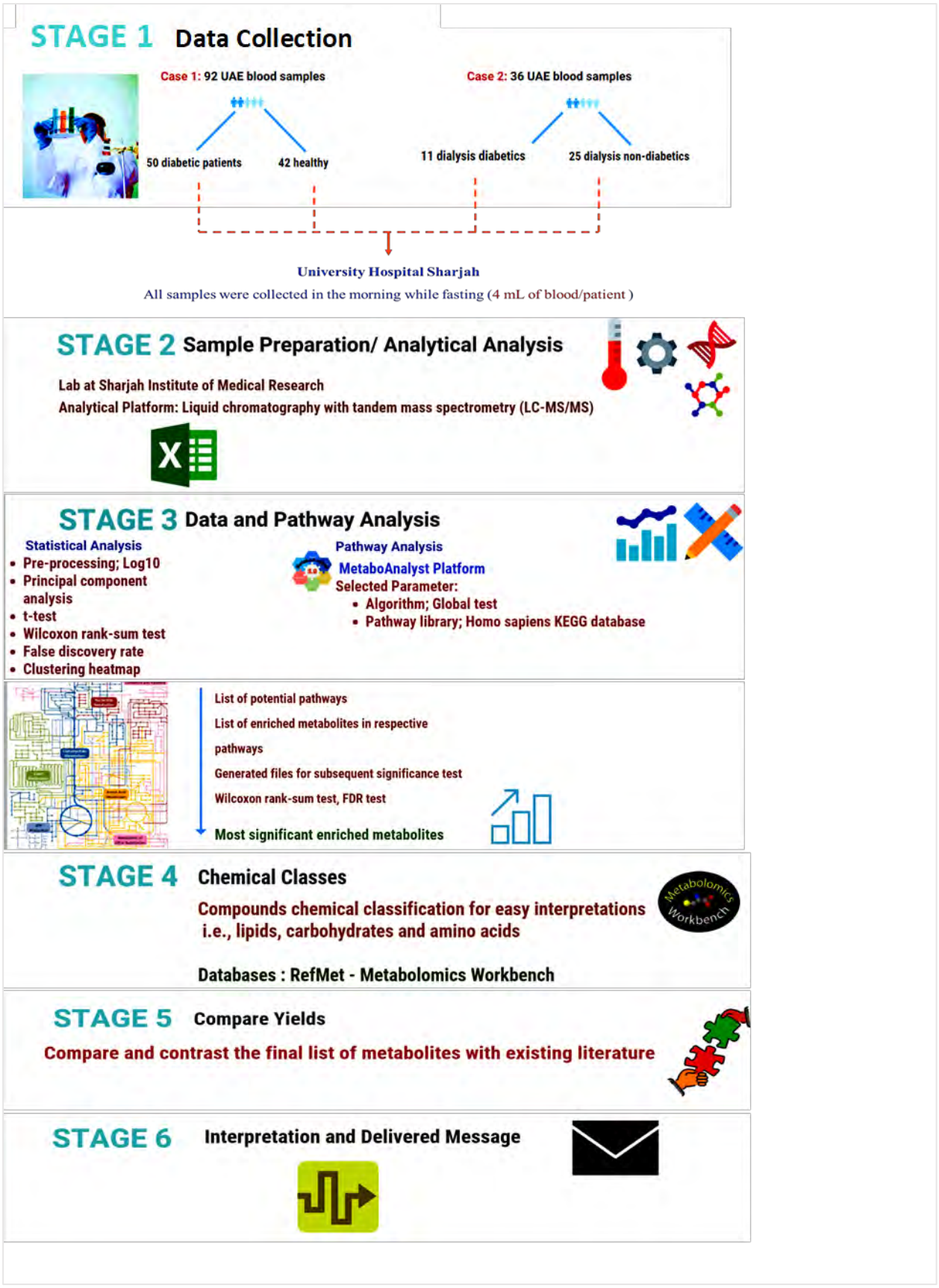


Figure 4-1: Sequential lists of methodological steps in the study.

4.3 Sample Preparation

A total of 4 mL of blood was collected from each subject into a sterile container. The samples were stored immediately at –80 ºC for long-term storage until further metabolomics analysis. All samples were collected daily (between 8 and 10 am) while fasting. An aliquot of plasma sample into a microcentrifuge tube and add cold methanol into the sample at 3:1 v/v (i.e., 30 µL sample, add 90 µL cold methanol) vortex and allow to sit in –20ºC for two hrs. Next, centrifuge the samples at 20,817 x g for 15 min at 4ºC. Then, transfer the supernatant to a new microcentrifuge tube. Usually, transfer three times the original sample volume (i.e., for 30 µL sample, add 90 µL cold methanol, then transfer 90 µL supernatant). Dry down the sample using Speed vac at 30 – 40°C. Store the dried sample in a –80ºC freezer for further use or dissolve it in solvent for LCMS analysis. Dissolve samples preferably in the starting solvent (0.1% formic acid) where volume is three times the original plasma volume. For example, when 30 µL serum/plasma has been used, dissolve the supernatant in 90 µL 0.1% formic acid. Place the vials in the autosampler.

4.4 Profiling Techniques and Analytical Measurement

The four main technologies used in the drug development field are NMR, the combination of LC-MS, its evolvement called UPLC-MS, and GC-MS. These different platforms do not compete, as none of them can conduct a complete detection and quantification of all metabolites set for a targeted biological sample. Accordingly, the optimum metabolomic experiments utilize various technology platforms [37], [235, 236]. Furthermore, MS and NMR provide complementary information; therefore, there is an appeal in merging NMR and MS techniques for disease research [237]. For example, combining MS and NMR enhances metabolite annotation and detection [236]. Therefore, combining several analytical sources is vital to the future of metabolomics research. Table 4-1 exhibits the main differences between NMR and MS platforms.

Table 4-1: Key distinctions between NMR and MS.

Feature/key point	NMR	MS	Ref.
Metabolite coverage/number	Few hundreds	Hundreds to thousands	[238], [235]
Sensitivity	Low	High	
Robustness	Extremely good	Very good	
Reproducibility	High	Lower than NMR	

Feature/key point	NMR	MS	Ref.
Apparatus cost	Expensive	LC: expensive, GC: non-expensive	
Resolution	Low	High	
Type of metabolite	Amino acids, Polar/nonpolar metabolites, Sugars, Volatile liquids, Large metabolites.	Amino acids, Fatty acids, Polar/ Nonpolar metabolites, Organic acids, Steroids, Volatile/thermally stable metabolites, Amino acids, Medium-to-high lipophilicity, Nucleosides and nucleotides, Carbohydrates, Esters.	

The robustness and availability of the MS technique in SIMR allowed us to identify and characterize molecules in our data. The MS output data signal is m/z spectrum. Figure 4-2 shows an MS spectrum output. The molecules analyzed by mass spectrum are charged; they are ions with a range of charges: +1, +2, and so on. The relative intensity or relative abundance are represented in the vertical axis; the two terms are used interchangeably.

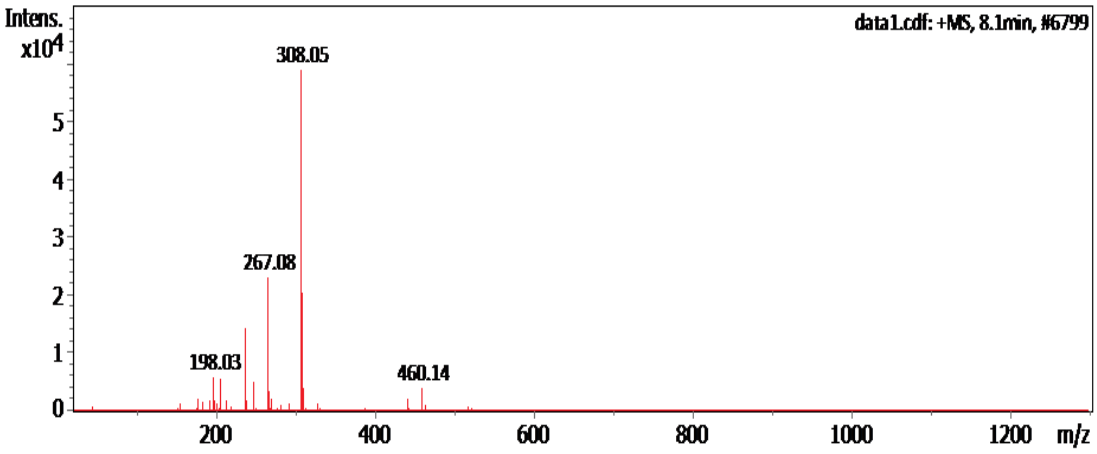


Figure 4-2: MS spectrum output.

TimsTOF Mass Spectrometer (BRUKER, Germany) and Metaboscape software version 4 in SIMR were employed to separate and detect the cell metabolites. It was equipped with a trapped quadrupole time-of-flight mass spectrometer and composed of Solvent delivery systems pump (ELUTE UHPLC Pump HPG 1300), Autosampler (ELUTE UHPLC), Thermostat column compartment (ELUTE UHPLC), Computer System, Windows 10 Enterprise 2016 LTSB, Data Management Software, Bruker Compass HyStar 5.0 SR1 Patch1 (5.0.37.1), Compass 3.1 for otofSeries, otofControl Version 6.0. Metabolites were analyzed in auto MS/MS positive scan mode within the range of 20-1300 m/z utilizing electrospray ionization (ESI). The ESI source was 10 L/min, and the drying temperature was equal to 220℃. The capillary voltage of the ESI was 4500 V with 2.2 bar nebulizer pressure. The collision energy was set at 7 eV and

end Plate Offset as 500 V. A HAMILTON ® Intensity Solo 2 C18 column (100 um x 2.1 mm x1.8 µm) was utilized the separation of metabolites. And Sodium Formate was used as a calibrant for the external calibration step. Solvent A (Water + 0.1% FA) and solvent B (Acetonitrile + 0.1% FA) were used in gradient elution mode for metabolite analysis. Metabolites were analyzed in auto MS/MS positive scan mode within the range of 20-1,300 m/z utilizing electrospray ionization (ESI). The ESI source with dry nitrogen gas was 10 l/min, and the drying temperature was equal to 220℃. The capillary voltage of the ESI was 4,500 V with 2.2 bar nebulizer pressure. For MS2 acquisition, the collision energy was set at 20 eV and end Plate Offset as 500 V. A Hamilton® Intensity Solo 2 C18 column (100 mm x 2.1 mm x 1.8 µm) was utilized to separate metabolites, and sodium formate was used as a calibrant for external calibration step. For metabolite analysis, solvent A (Water + 0.1% FA) and solvent B (Acetonitrile + 0.1% FA) were used in gradient elution mode. The gradient program used a flow rate of 0.250 ml/min with 99A:1.0B from 0.00-2.00 min, 99A:1.0B to 1.0A:99B from 2.00-17.00 min, 1.0A:99B from 17.00-20.00 min, 1.0A:99B to 99A:1.0B from 20.00-20.10 min, flow rate of 0.350 ml/min with 99A:1.0B from 20.10-28.50 min, flow rate of 0.250 ml/min, with 99A:1.0B from 28.50-30 min giving a total run time of 30 min with a maximum pressure of 14993 pounds per square inch (PSI). The autosampler temperature was set at 8℃ and the column oven temperature at 35℃. A total volume of 10 µl was injected into the QTOF MS. The flow rate was set as (0.250-0.350 mL/min) for 30 min in gradient mode with a maximum pressure of 14993 psi. The elute autosampler temperature was set at 8℃, and the column oven temperature was at 35℃. And a total volume of 10 µL was injected into the QTOF MS. LC total ion chromatograms (TIC), and fragmentation patterns of the metabolites were identified by MetaboScape® version 4.0 (Bruker-Daltonics) and MS/MS library search based on the Bruker HMDB Metabolite Library 2.0 (Bruker Daltonics). The latter library provides more than 6000 MS/MS spectra for more than 800 compounds selected from the HMDB [45]. Data processing. Processing and statistical analysis were performed using MetaboScape® 4.0 software (Bruker Daltonics). Bucketing in T-ReX 2D/3D workflow, the parameters set for molecular feature detection were as follows: minimum intensity threshold equal to 1,000 counts along with minimum peak length of 7 spectra for peak detection, using peak area for feature quantitation. The mass recalibration was done within a 0-0.3 min retention time range. Only those features present in at least 3 of 12 samples (per cell type) were considered. On the other hand, the MS/MS import

method was set to be done by average. The parameters for data bucketing were assigned as follows: Retention time range started at 0.3 min and ended at 25 min, while mass range started at 50 m/z and ended at 1,000 m/z. Each sample was run in duplicate LC-MS/MS analysis as described above.

Bland Altman plots (Figure 4-3) were used to compare and estimate bias and agreement between the duplicate analytical measurements. The Bland Altman plot analysis is a simple way to evaluate a bias between the mean differences, and to estimate an agreement interval, within which 95% of the differences of the second method, compared to the first one fall. Exploring the agreement analysis for the whole data, we can say that biologically the agreement interval is not wide and sufficiently narrow for our purpose.

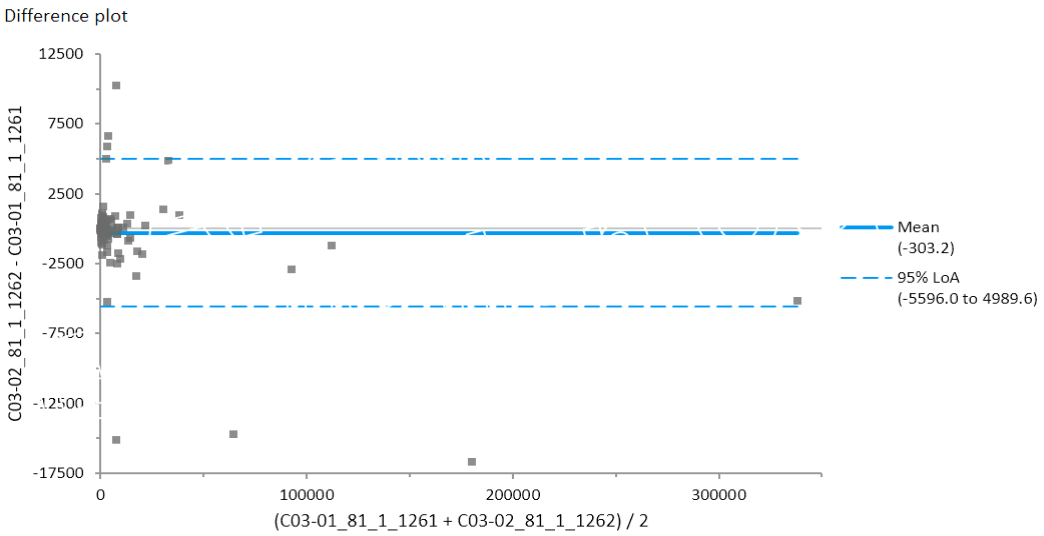


Figure 4-3: Plot of differences between measurement A and measurement B vs. the mean of the two measurements for sample 3.

4.5 Output Data Format

Generally, the chromatogram representation of raw signals is a graphical description of separated eluents managed to detect compounds and define their relative concentrations in the spectra. Due to the heterogeneity of data format, it is challenging for researchers to manipulate and share such data. Therefore, numerous software programs for converting proprietary raw data file forms such as .CSV, .TXT, .mzXML, or .netCDF, into a universal format have been employed..

A mass spectrometer output file format mzML was developed [239]. The mzML format is an open, based on extensible markup language (XML) format for mass spectrometer

output files to create a distinct open format sponsored by each software. The format incorporates the best aspects from pre-existing open formats and has further help for chromatograms and other desirable characteristics.

Output data are typically structured in matrices with rows representing samples and columns corresponding to features. There are different standard input data formats: metabolite concentration, spectra bins/peak table, NMR/MS peak lists, and GC/LC-MS raw spectra. MetaboAnalyst use data in several types of tabular format such as textual tab-delimited TXT (.txt) or comma-separated value (.csv). Also, MetaboAnalyst receives zipped files of MS peak lists or MS spectra that must be in mzXML [240], mzDATA [241], or NetCDF [242] open data format [55]. After converting raw data to an appropriate format, i.e., (.csv) format, we can further analyze them.

4.6 Statistical Data Analysis

R software version 4.0.5 was used for the statistical analysis [204]. Data was analyzed in a duplicate technique. Data cleaning excluded concentration values that were missing or below the detection limit. Then, each sample was averaged. Data standardization and normalization were performed through Logarithmic transformation following standard normalization techniques found in related literature.

The development of biomarkers into diagnostic or prognostic tests can be categorized into three broad phases: discovery, performance evaluation, and impact determination when added to existing clinical measures. Each stage requires a unique study design and statistical considerations to accomplish research objectives accurately. The necessary statistical methodology for assessing biomarker performance differs from the classic methods used in epidemiology or therapeutic research. However, the biomarker discovery stage focus on measures of association.

Initially, we run a non-parametric Kruskal Wallis Test to tell whether the overall comparison is significant, and post-hoc analyses usually follow it to identify which two levels are different. Differential metabolites between the different groups of patients were identified using PCA and Wilcoxon rank-sum test (known as Mann-Whitney U-test). Wilcoxon rank-sum test does not assume our data have a known distribution. In addition, the False Discovery Rate (FDR) method was applied to adjust for the multiple comparisons problem. A 0.05 significance level was assumed throughout the analysis, and adjusted p-values through FDR less than 0.05 were assumed significant.

4.7 Pathway Analysis

MetaboAnalyst 5.0 platform [211] introduces the PA module for pathway enrichment and topological analysis. Pathway enrichment analysis computes a single P-value for each metabolic pathway (a group of functional-associated metabolites) instead of the t-test, which determines the statistical significance of the difference between individual metabolites. Pathway topology analysis utilizes graph theory to evaluate a given experimentally identified metabolite's importance in a pre-defined metabolic pathway. Measurements were computed using centrality, a standard metric used in graph theory to estimate the relative importance of individual nodes to the overall network. A "pathway impact score" was then computed as the sum of the important measures of identified metabolites divided by the total sum of the important measures of all the identified and unidentified metabolites in the pathway. The pathway impact score represents an objective estimate of the importance of a given pathway relative to a global metabolic network. First, we uploaded the data input files to MetaboAnalyst 5.0 web server, where data pre-processing, such as normalization and scaling, and metabolic PA were performed. Then, we specified PA parameters as follows: (1) global test method [243] performed enrichment analysis, (2) Relative Betweenness was used to measure centrality, and (3) 80 human metabolic pathways (Homo sapiens) in the KEGG database were employed as reference metabolic pathways. The choice of FDR value for selecting the best number of pathways is arbitrary. Commonly, the FDR value is identified based on the p-value and impact score value. The nodes with the most significant p-value (more dense color) and the nodes with the higher impact score values (bigger size) are chosen. Figure 4-2 lists sequential methodological steps in our work.

Chapter 5. Metabolomics Profile for T2DM Emirati Population versus Healthy: Untargeted Approach

5.1 Introduction

Metabolomics has great potential as a decision-making tool offering valuable information on the physiological state. The minor change in the expression levels of genes or proteins causes an overt variation in the level of metabolites. In addition, chronic diseases occur from the impact of multi factors, such as genetics, lifestyle, and environment. Therefore, to compare the metabolite concentration levels in phenotypically recognized populations, e.g., diseased and control subjects. It might support identifying etiological pathways and biological processes that have significantly changed across two different biological states. However, the high dimensionality of the metabolomics observations often complicates the interpretation of the findings. Pathway analysis is a standard method of studying gene expression data. Focusing only on the over-represented subsets in the outcomes of metabolomics assays can substantially reduce dimensionality. This approach, known as over-representation, or enrichment analysis, has become one of the standard tools for interpreting high-throughput metabolomics observations with the primary purpose of dimensionality reduction [244].

Studies about the metabolomic profile of T2DM from the Middle Eastern populations are still in their early stages. Although metabolomics reports identified several metabolites whose levels are related to dysglycaemia and T2DM, further studies need to be conducted in biomarker discovery—specifically, studies on DM in the MENA region. Examining common metabolite biomarkers in various studies and a biochemical relationship with the disease will identify those with higher potential [68]. Therefore, this study explores the metabolomic profile of T2DM and non-T2DM UAE citizens to uncover the potential novel biomarkers in this population.

5.2 Materials and Methods

5.2.1 Patients

A case-control analysis was conducted on blood specimens collected from Emirati citizens. Ninety-two subjects (50 T2DM and 42 non-T2DM) were collected from University Hospital Sharjah. All samples were collected in the morning while fasting.

5.2.2 Sample collection, preparation, and analytical analysis

A total of 4 mL of blood was collected from each patient after overnight fasting into a sterile container. All samples were assembled at roughly the same time each day (between 8 and 10 am every day). The samples preparation method was described previously in the Methods chapter.

TimsTOF Mass Spectrometer (BRUKER, Germany) and MetaboScape software version 4 (Brucker) were employed to separate and detect the cell metabolites. Detailed explanation about LC-MSMS analytical techniques is also found in the Methods chapter.

5.2.3 Statistical and pathway analysis

We attained statistical analysis using R software version 4.0.5 [220]. First, too many zeroes or missing values will cause difficulties for downstream analysis. Therefore, the default method replaces all the missing and zero values with small ones; zeros are replaced with 2. This approach assumes that most missing values are caused by low abundance metabolites (i.e., below the detection limit). In addition, since zero values may cause data normalization problems (i.e., log), they are also replaced with this small value. Then, each sample was averaged as each sample was duplicated. Data transformation applies a mathematical transformation on individual values themselves. The logarithmic transformation (base 10) technique is used.

Differential metabolites between the 50 T2DM and 42 non-T2DM patients were detected using PCA and Wilcoxon rank-sum test (known as Mann-Whitney U-test). They provide a preliminary overview of potentially significant features in discriminating the conditions under study. In addition, the FDR method was applied to adjust for the multiple comparisons problem. A 0.05 significance level was assumed throughout the analysis, and adjusted p-values through FDR less than 0.05 were considered significant.

MetaboAnalyst 5.0 platform [211] introduces the Pathway Analysis module for pathway enrichment and topological analysis. Pathway enrichment analysis computes a single P-value for each metabolic pathway (a group of functional-associated metabolites) instead of the t-test, which determines the statistical significance of the difference between individual metabolites. Pathway topology analysis utilizes graph theory to evaluate a given experimentally identified metabolite's importance in a pre-

defined metabolic pathway. Measurements were computed using centrality, a standard metric used in graph theory to estimate the relative importance of individual nodes to the overall network. A "pathway impact score" was then computed as the sum of the important measures of identified metabolites divided by the total sum of the important measures of all the identified and unidentified metabolites in the pathway. The pathway impact score represents an objective estimate of the importance of a given pathway relative to a global metabolic network.

First, we uploaded the data input files to MetaboAnalyst 5.0 web server [211], where data pre-processing, such as normalization and scaling, and metabolic pathway analysis were performed. Then, we specified pathway analysis parameters as follows: (1) global test method [251] performed enrichment analysis, (2) Relative Betweenness was used to measure centrality, and (3) 80 human metabolic pathways (Homo sapiens) in the KEGG database were employed as reference metabolic pathways. The Figure shows part of the matched pathways and their enriched metabolites according to the p values from the pathway enrichment analysis and pathway impact values from the pathway topology analysis for nondiabetics vs. uncontrolled diabetics groups. We selected potential pathways based on the arbitrary FDR cut-off values (fdrcut-off) exhibited in the results section. The choice of fdrcut-off since was no considerable increase in significantly enriched pathways when the fdrcut-off was increased beyond the preferred values. Subsequently, the matched metabolites in each pathway are documented for further analysis. Finally, the resulting excel file containing all matched metabolites from the selected pathways is analyzed using the Wilcoxon rank test to filter the most significant metabolite between different groups. As shown in Figure, the compound colors within the pathway are as the following: light blue means those metabolites are not in our data and are used as background for enrichment analysis; grey means the metabolite is not in our data and is also excluded from enrichment analysis; other colors varying from yellow to red means the metabolites are in the data with different levels of significance. Each potential pathway should be visited to record matching metabolites to combine our final document. Figure 5-1 visualizes an example for pathway analysis conducted in MetaboAnalyst.While collecting matching metabolites, one of the main concerns is using different names for the same metabolites across databases and various studies. Therefore, we used RefMet, a reference list of metabolite names [202]. RefMet is essential for comparing and contrasting metabolite

data across different experiments and studies. Finally, the KEGG pathway is a collection of manually drawn pathway maps representing molecular interaction, reaction, and relation networks. Figure 5-2 displays the KEGG Homo sapiens pathway map for Aminoacyl-tRNA biosynthesis (hsa00970). However, the very intricate nature of interaction and pathway signaling makes inter-pathway dependence the most critical challenge in pathway analysis up to now. The metabolome view figures showing all matched pathways according to the p values from the pathway enrichment analysis and pathway impact values from the pathway topology analysis are demonstrated in the results section.

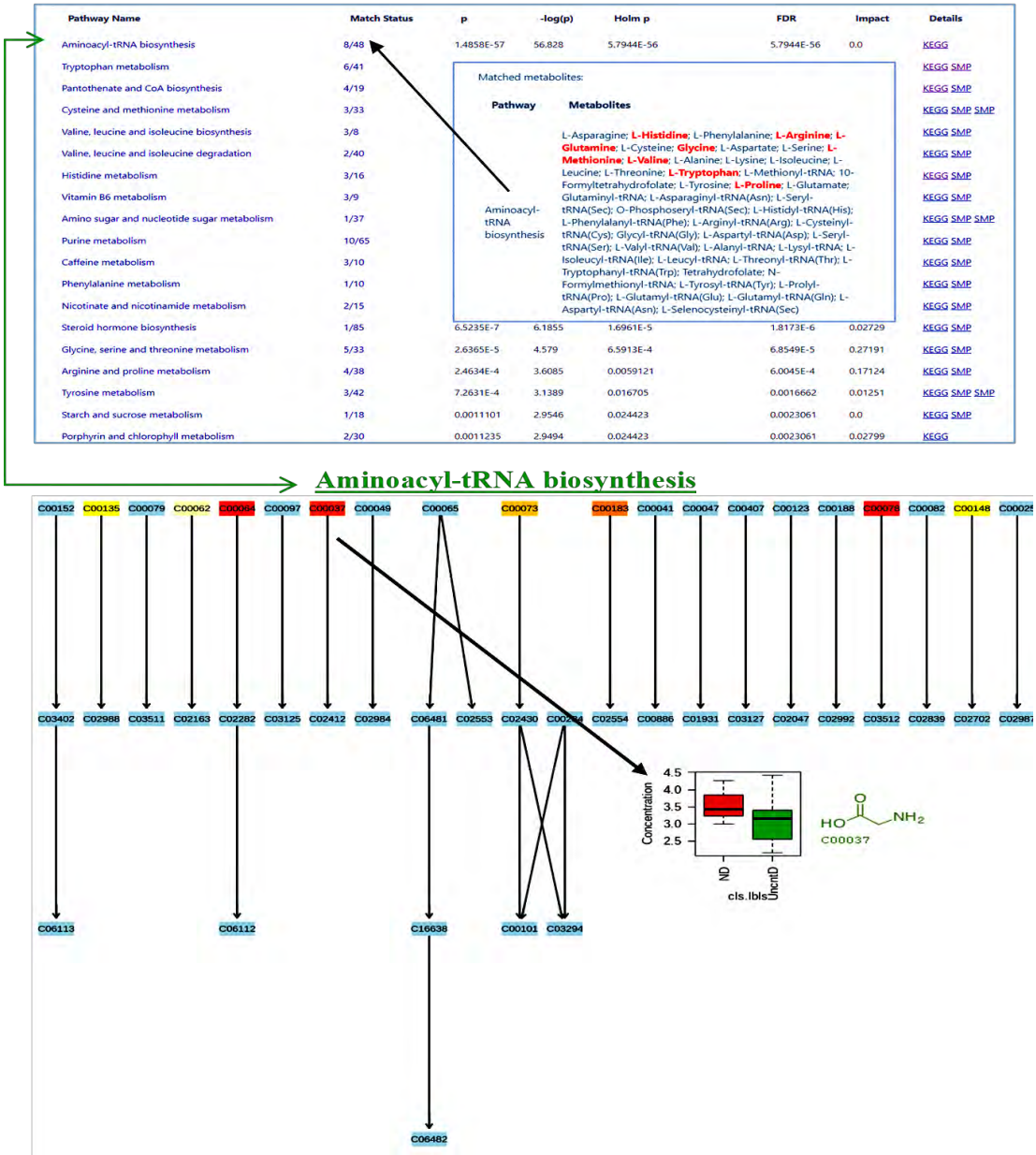


Figure 5-1: Graphical visualization for pathway analysis conducted in MetaboAnalyst, Aminoacyl-tRNA biosynthesis pathway is chosen as an example for methods explanation.

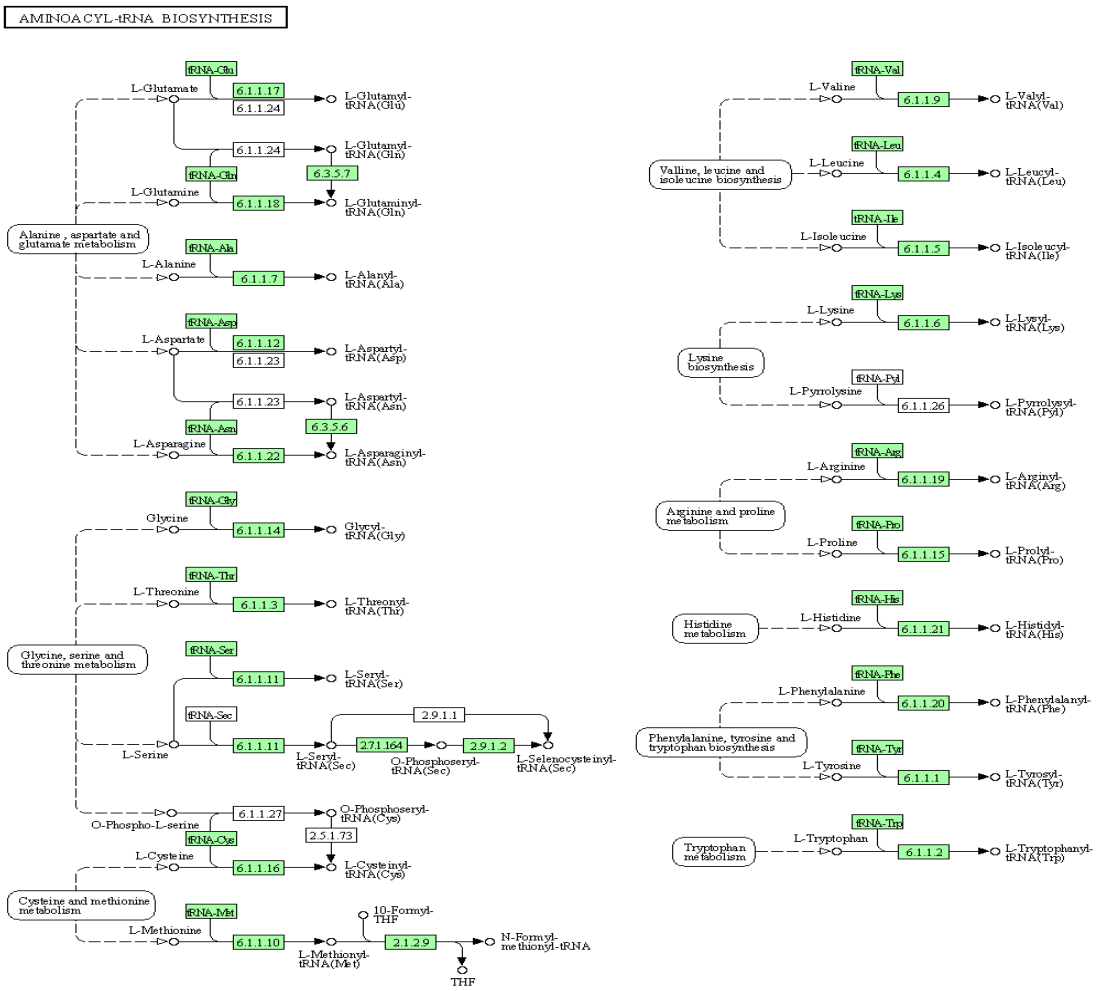


Figure 5-2: KEGG pathway map for Aminoacyl-tRNA biosynthesis (hsa00970).

5.3 Results

5.3.1 Clinical data of patients

Ninety-two subjects were voluntarily enrolled in this study, of which 50 are diagnosed with T2DM and are being treated at the University Hospital Sharjah. The other 42 subjects have no known T2DM status and refer to them as a non-T2DM group. In T2DM, 35 females aged between 23 and 86 (average: 65.5 ± 15.5 years; median: 71 years), and 15 males aged between 18 and 85 (average: 68.9 ± 17.4 years; median: 72 years). In non-T2DM patients, 31 females aged between 18 and 68 (average: 33.7 ± 12.3 years; median: 28 years), and 11 males aged between 25 and 88 (average: 42.6 ± 18.9 years; median: 36 years). The classification for patients is based on the clinically confirmed diabetic status according to WHO diagnostic criteria for diabetes (fasting plasma glucose ≥ 7.0mmol/l (126mg/dl) or 2–hrs. plasma glucose ≥ 11.1mmol/l (200mg/dl)). Patients' demographic data are presented in Table 5-1.

Table 5-1: Demographics characteristics of individuals with and without diabetes.
T2DM: diabetic patients, n: sample size, BMI: body mass index, SD: standard deviation, M: male, F: female.

Characteristics	T2DM **(n=50)**	Non-T2DM **(n=42)**
Age in years: mean (SD, range) BMI: mean(SD, range)	66.48 (15.9, 68) 28.76 (4.81, 27.095)	36.02(14.60, 70) 27.04 (5.56, 20.894)
Gender: (M%. F%))	(30%, 70%)	(26%, 74%)

5.3.2 Differential metabolite screening

Integration of LC-MS/MS technique and HMDB database [45] revealed 148 detected and identified metabolites. In addition, we used the MetaboAnalyst 5.0 platform to examine the patterns of these identified metabolites. The top 50 metabolites based on the differences in averages between T2DM and non-T2DM groups are displayed as a heatmap in Figure 5-3. Columns represent samples, rows represent metabolites, and the relative content of the metabolites is displayed by color. Heatmap in Figure 5-3 indicates apparent differences in the concentration of the metabolites among the two groups. Examples include Salicyluric acid, 4-Pyridoxic acid, 2-Pyrrolidinone, and Indoleacetic acid.

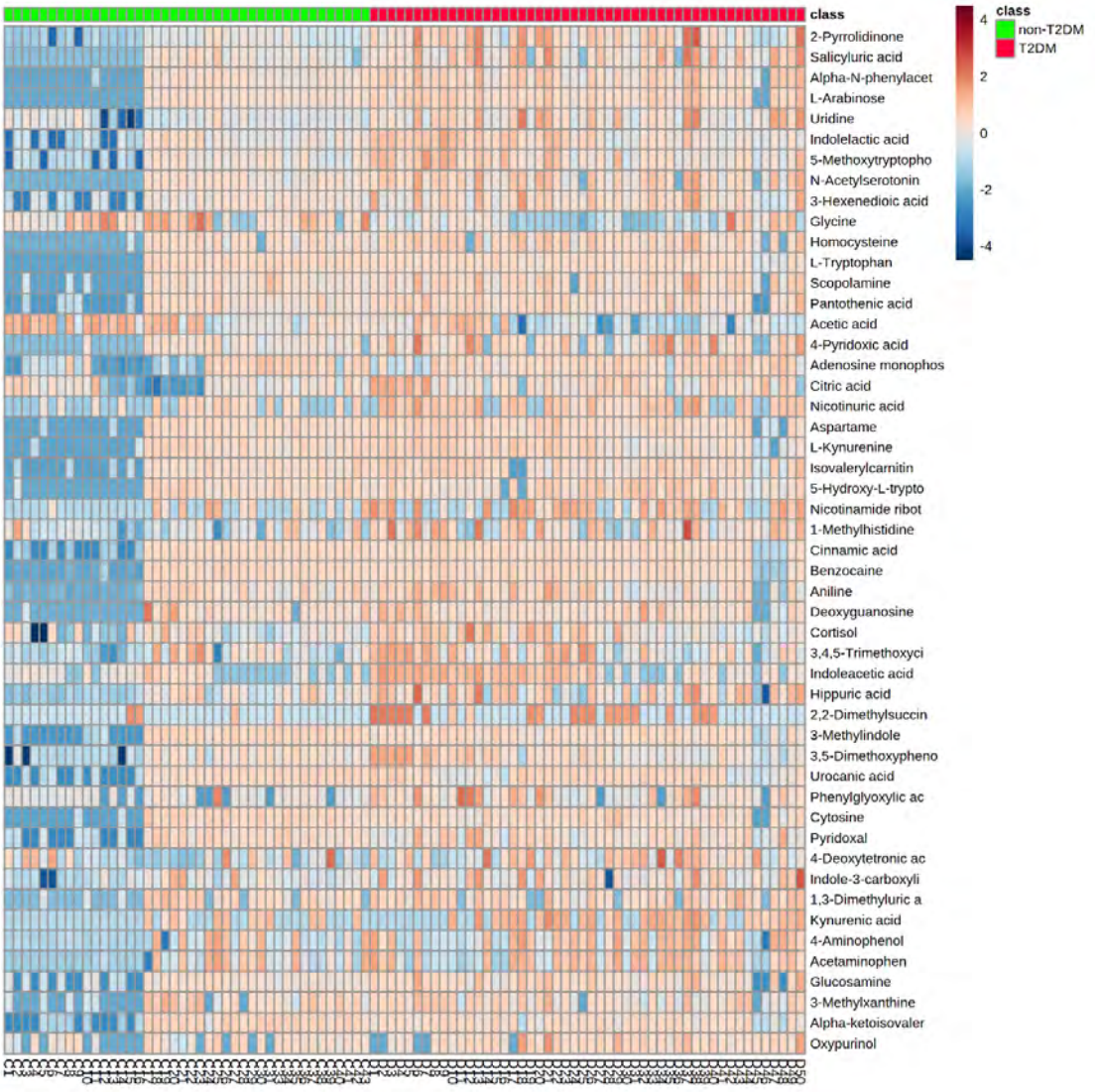


Figure 5-3: Heatmap of the 50 selected metabolites among the T2DM and non-T2DM patients.

5.3.3 Multivariate statistical analysis

First, statistical analysis was performed based on clinically confirmed diabetic status. The plot of the top two principal components following the PCA analysis of the 148 identified metabolites grouped by clinically confirmed diabetic status is shown in Figure 5-4 A. Figure 5-4 A depicts the blood components of the T2DM group, and the non-T2DM group has apparent clustering, particularly for the non-T2DM group. However, the available HbA1c data and BMI values indicate new potential groups. Therefore, the data were explored based on the recent HbA1c and BMI values indicated in Figure 5-4 B. PCA plot in Figure 5-4 B shows three main groups as follows: (1) non-diabetics (ND) patients, (2) uncontrolled diabetics (Uncontrolled D), and (3) controlled diabetics and prediabetics (Pre/controlled D). In all subsequent analyses, we assume three groups of subjects as indicated in Figure 5-4 B.

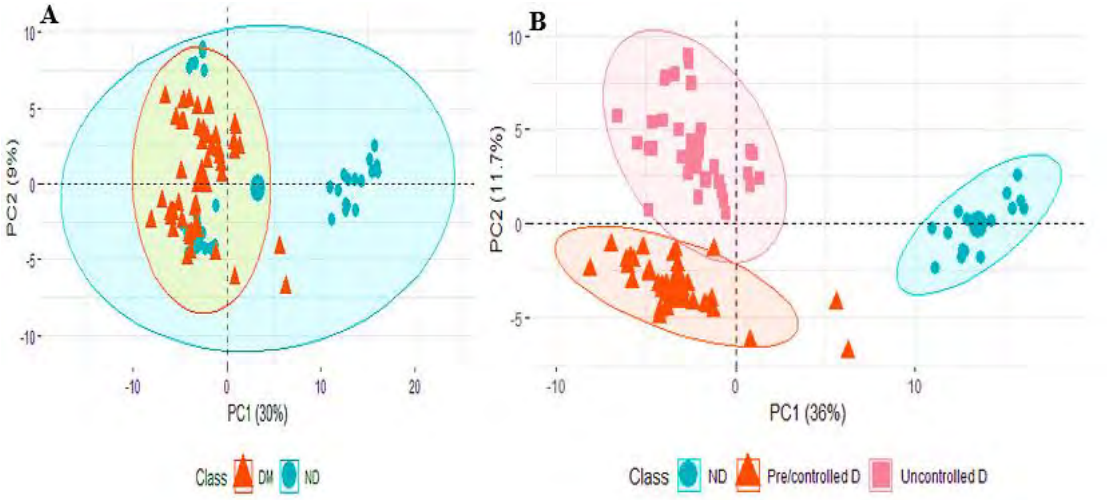


Figure 5-4: Plots of PCA scores. (A) PCA plot based on clinically confirmed diabetic status, (B) PCA plot shows new groups based on most recent HbA1c values and BMI.

5.3.4 Differential metabolite analysis

Following the grouping of subjects in three groups shown in Figure 5-4 B, three scenarios were analyzed to search for differential metabolites accounting for the three possible groupwise comparisons: (1) non-diabetics vs. uncontrolled diabetics, (2) non-diabetics vs. prediabetics, and controlled diabetics, and (3) uncontrolled diabetics vs. prediabetics and controlled diabetics. Heatmaps were generated for the three pairwise comparisons to visualize the metabolomics data based on the t-test. Figure 5-5 shows the top 50 metabolites between non-diabetics vs. uncontrolled diabetics. Figure 5-6 depicts the principal 50 metabolites between non-diabetics vs. prediabetics and controlled diabetics. Figure 5-7 exhibits the highest 50 metabolites between uncontrolled diabetics and prediabetics and controlled diabetics. The initial visualization inspection based on Figures 5-5, 5-6, and 5-7 indicates an apparent variability between metabolites levels among the identified groups. Therefore, further analysis is required to better understand the metabolic changes in such groups. The non-parametric Wilcoxon rank-sum test analyzed the differential metabolites among the three previously identified groups. Then, FDR adjusted P-values were attained. In non-diabetics and uncontrolled diabetics groups, a total of 95 significant metabolites listed in Table 5-2 were observed. In non-diabetics vs. prediabetics and controlled diabetics groups, 117 significant metabolites recorded in Table 5-3 were examined. In uncontrolled diabetics vs. prediabetics and controlled diabetics groups, 50 significant metabolites listed in Table 5-4 were spotted. MetabolomicsWorkbench RefMet was used to match each metabolite with its corresponding subclass [194].

MetabolomicsWorkbench RefMet is a standardized reference terminology for metabolomics. RefMet is essential for comparing and contrasting metabolite data across different experiments and studies.

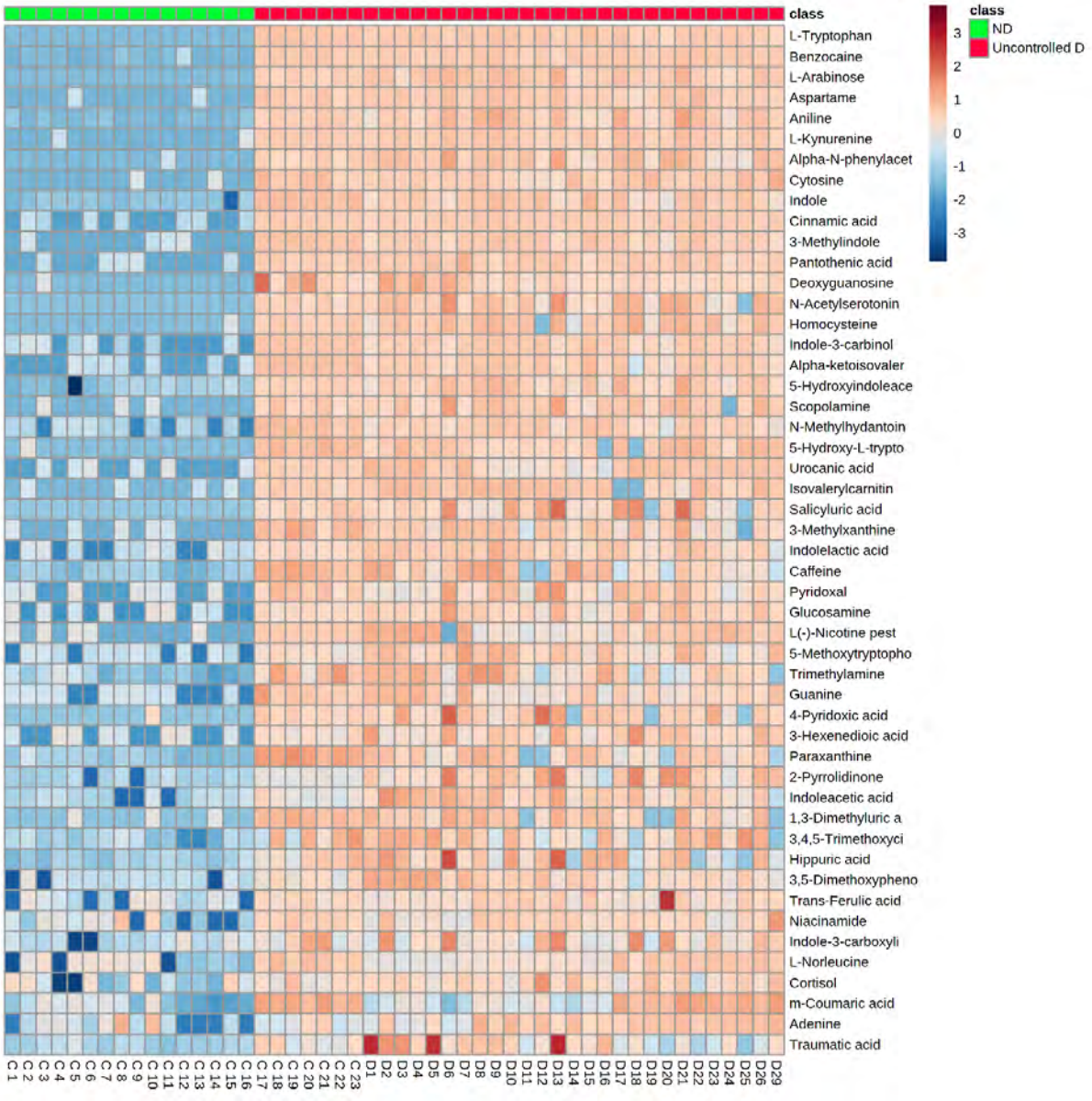


Figure 5-5: Heatmap of the 50 selected (t-test) metabolites among the ND and Uncontrolled D.

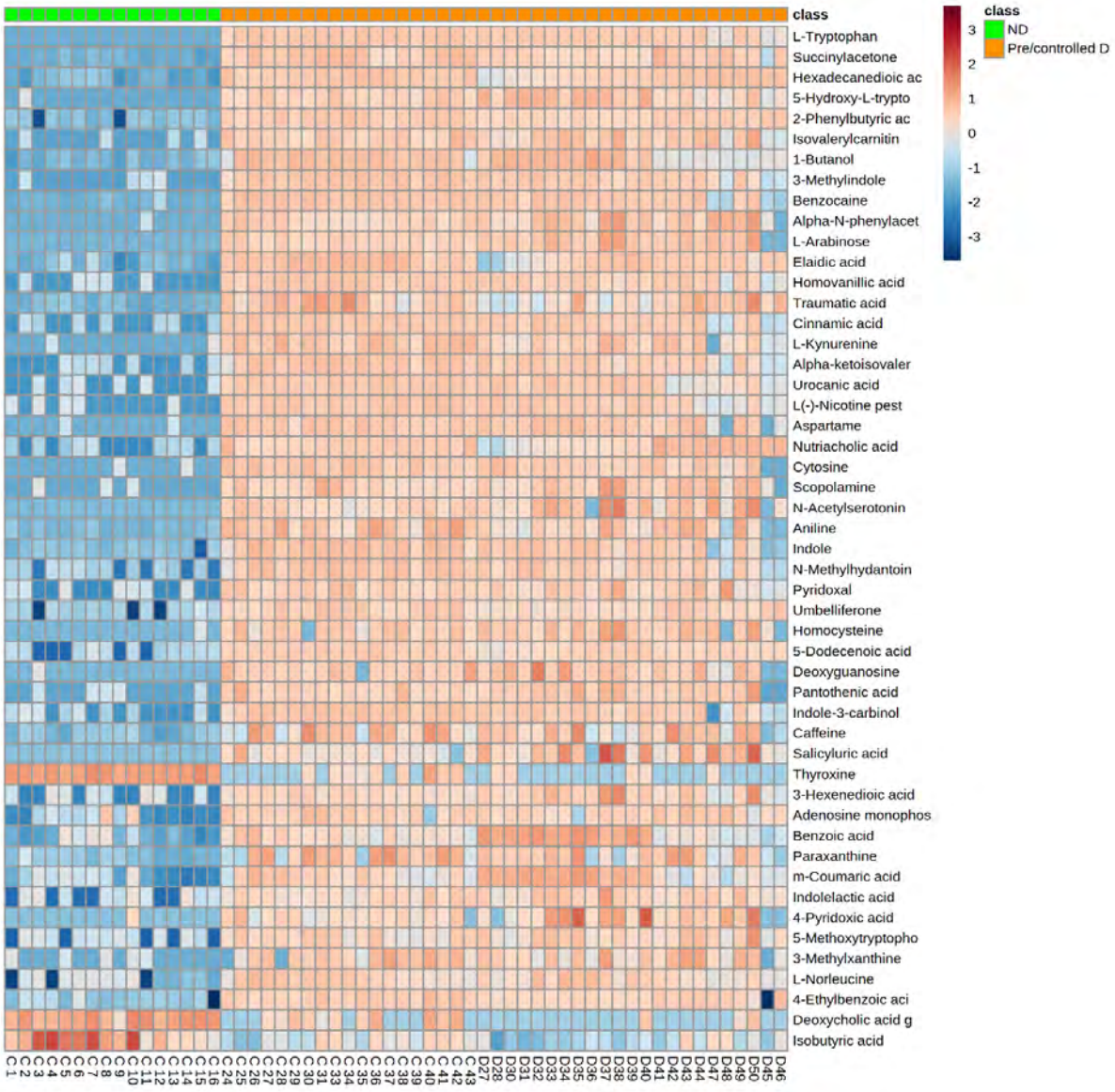


Figure 5-6: Heatmap of the 50 selected metabolites (t-test) among the ND and Pre/controlled D.

Figure 5-7: Heatmap of the 48 significant metabolites (t-test) among the Uncontrolled D and Pre/controlled D.

Table 5-2: List of significant metabolites between non-diabetics and uncontrolled diabetics (Wilcoxon rank-sum test).

Metabolite	FDR	Subclass	Metabolite	FDR	Subclass
5-Hydroxyindoleacetic acid	<0.001	Indole-3-acetic acid derivatives	**Indole-3-carboxylic acid**	<0.001	Indolecarboxylic acids
Aniline	<0.001	Anilines	**4-Pyridoxic acid**	<0.001	Pyridinecarboxylic acids
Benzocaine	<0.001	Benzoic acids	**Succinylacetone**	<0.001	Medium-chain keto acids
Acetone	<0.001	Ketones	**4-Aminophenol**	<0.001	Aminophenols
Trimethylamine	<0.001	Tertiary amines	**1,3-Dimethyluric acid**	<0.001	Xanthines
3,4,5-Trimethoxycinnamic acid	<0.001	Coumaric acids	**Traumatic acid**	<0.001	Dicarboxylic acids
Hippuric acid	<0.001	Hippuric acids	**Pyridoxal 5'-phosphate**	<0.001	Pyridoxals
Paraxanthine	<0.001	Xanthines	**Acetaminophen**	<0.001	Aminophenols
L-Arabinose	<0.001	Monosaccharides	**Urea**	<0.001	Isoureas
L-Tryptophan	<0.001	Amino acids	**m-Coumaric acid**	<0.001	Hydroxycinnamic acids
Alpha-N-phenylacetyl-L-glutamine	<0.001	Amino acids	**Guanidoacetic acid**	<0.001	Amino acids
Deoxyguanosine	<0.001	Purine deoxyribonucleosides	**Niacinamide**	<0.001	Nicotinamides
Aspartame	<0.001	Peptides	**L-Norleucine**	<0.001	Amino FA
Cytosine	<0.001	Pyrimidones	**Cortisol**	<0.001	C21 steroids
L-Kynurenine	<0.001	Butyrophenones	**o-Tyrosine**	<0.001	Amino acids
3-Methylindole	<0.001	Indoles	**Paracetamol sulfate**	<0.001	Phenylsulfates
Pantothenic acid	<0.001	Amino acids	**Nicotinuric acid**	<0.001	Amino acids
Cinnamic acid	<0.001	Cinnamic acids	**PC(18:1(9Z)/18:1(9Z))**	<0.001	PC
Indole-3-carbinol	<0.001	Indoles	**Thyroxine**	<0.001	Diarylethers
Guanine	<0.001	Hypoxanthines	**Benzoic acid**	<0.001	Benzoic acids
5-Methoxytryptophol	<0.001	Indoles	**Phenylglyoxylic acid**	<0.001	Phenylacetic acids
N-Methylhydantoin	<0.001	Hydantoins	**Adenine**	<0.001	6-aminopurines
Indole	<0.001	Indoles	**Nicotinamide ribotide**	<0.001	Nicotinamide nucleotides
Alpha-ketoisovaleric acid	<0.001	Branched FA	**ADP**	<0.001	Purine rNDP
Urocanic acid	<0.001	Imidazolyl carboxylic acids	**Dehydroascorbic acid**	<0.001	Gamma butyrolactones
2-Pyrrolidinone	<0.001	Pyrrolidine-2-ones	**Citrulline**	<0.001	Amino acids
Pyridoxal	<0.001	Pyridoxals	**DL-2-aminooctanoic acid**	<0.001	Amino acids
3-Hexenedioic acid	<0.001	Dicarboxylic acids	**Uridine**	<0.001	Pyrimidine ribonucleosides
Glucosamine	<0.001	Amino sugars	**Hypoxanthine**	0.001	Hypoxanthines
Indolelactic acid	<0.001	Indolyl carboxylic acids	**Adenosine monophosphate**	0.001	Purine rNMP

Metabolite	FDR	Subclass	Metabolite	FDR	Subclass
N-Acetylserotonin	<0.001	Serotonins	**Kynurenic acid**	0.001	Quinoline carboxylic acids
Trans-Ferulic acid	<0.001	Hydroxycinnamic acids	**Trehalose**	0.001	Disaccharides
Homocysteine	<0.001	Amino acids	**N6-Acetyl-L-lysine**	0.002	
Indoleacetic acid	<0.001	Indole-3-acetic acid derivatives	**5-Aminolevulinic acid**	0.003	Amino FA
Scopolamine	<0.001		**Oxypurinol**	0.003	Xanthines
Salicyluric acid	<0.001	Hippuric acids	**3-Indolepropionic acid**	0.003	Indolyl carboxylic acids
3-Methylxanthine	<0.001	Xanthines	**Homo-L-arginine**	0.005	Amino acids
5-Hydroxy-L-tryptophan	<0.001	Amino acids	**Acetic acid**	0.006	Saturated FA
3,5-Dimethoxyphenol	<0.001		**L-Glutamine**	0.006	Amino acids
L(-)-Nicotine pestanal	<0.001		**Ribothymidine**	0.006	Pyrimidine ribonucleosides
Isovalerylcarnitine	<0.001	Acyl carnitines	**Acetaminophen glucuronide**	0.007	Sugar acids
Caffeine	<0.001	Xanthines	**Isovalerylglycine**	0.007	Amino acids
Homoveratric acid	0.012	Phenylacetic acids	**Uric acid**	0.030	Xanthines
Pyroglutamic acid	0.012	Pyrroline carboxylic acids	**Phenylpropiolic acid**	0.044	Benzenes
1-Methylhistidine	0.020	Amino acids	**Creatinine**	0.029	Imidazolines
Glycine	0.021	Amino acids	**Gallic acid**	0.030	Gallic acids
2,5-Furandicarboxylic acid	0.022		**Epinephrine**	0.028	Catechols
Protocatechuic acid	0.025	Hydroxybenzoic acids			

Table 5-3: List of significant metabolites between non-diabetics and prediabetics/controlled diabetics (Wilcoxon rank-sum test).

Metabolite	FDR	Subclass	Metabolite	FDR	Subclass
Hexadecanedioic acid	<0.001	Dicarboxylic acids	Urea	<0.001	Isoureas
Traumatic acid	<0.001	Dicarboxylic acids	Aniline	<0.001	Anilines
Benzocaine	<0.001	Benzoic acids	Aspartame	<0.001	Peptides
Elaidic acid	<0.001	Unsaturated FA	Cytosine	<0.001	Pyrimidones
Paraxanthine	<0.001	Xanthines	4-Ethylbenzoic acid	<0.001	Benzoic acids
Protocatechuic acid	<0.001	Hydroxybenzoic acids	Pyroglutamic acid	<0.001	Pyrroline carboxylic acids
m-Coumaric acid	<0.001	Hydroxycinnamic acids	Indole	<0.001	Indoles
Thyroxine	<0.001	Diarylethers	Pantothenic acid	<0.001	Amino acids
Hippuric acid	<0.001	Hippuric acids	Indole-3-carbinol	<0.001	Indoles
L-Tryptophan	<0.001	Amino acids	Glycocholic acid	<0.001	C24 bile acids
5-Hydroxy-L-tryptophan	<0.001	Amino acids	Homocysteine	<0.001	Amino acids
Isovalerylcarnitine	<0.001	Acyl carnitines	3,5-Dimethoxyphenol	<0.001	
Pyridoxal 5'-phosphate	<0.001	Pyridoxals	Deoxyguanosine	<0.001	Purine deoxyribonucleosides
5-Dodecenoic acid	<0.001	Unsaturated FA	o-Tyrosine	<0.001	Amino acids

Metabolite	FDR	Subclass	Metabolite	FDR	Subclass
Umbelliferone	<0.001	Hydroxycoumarins	Trans-Ferulic acid	<0.001	Hydroxycinnamic acids
1-Butanol	<0.001	Fatty alcohols	3-Methylxanthine	<0.001	Xanthines
2-Phenylbutyric acid	<0.001		Benzoic acid	<0.001	Benzoic acids
Succinylacetone	<0.001	Medium-chain keto acids	4-Aminophenol	<0.001	Aminophenols
5-Methoxytryptophol	<0.001	Indoles	Guanine	<0.001	Hypoxanthines
Urocanic acid	<0.001	Imidazolyl carboxylic acids	Glucosamine	<0.001	Amino sugars
Cinnamic acid	<0.001	Cinnamic acids	PC(18:1(9Z)/18:1(9Z))	<0.001	PC
Nutriacholic acid	<0.001		Trehalose	<0.001	Disaccharides
L-Norleucine	<0.001	Amino FA	Trimethylamine	<0.001	Tertiary amines
2-Pyrrolidinone	<0.001	Pyrrolidine-2-ones	3-Indolepropionic acid	<0.001	Indolyl carboxylic acids
Homovanillic acid	<0.001	Phenylacetic acids	DL-2-aminooctanoic acid	<0.001	Amino acids
Alpha-ketoisovaleric acid	<0.001	Branched FA	4-Pyridoxic acid	<0.001	Pyridinecarboxylic acids
3-Methylindole	<0.001	Indoles	Creatinine	<0.001	Imidazolines
Pyridoxal	<0.001	Pyridoxals	Adenosine monophosphate	<0.001	Purine rNMP
L(-)-Nicotine pestanal	<0.001		Dodecanedioic acid	<0.001	Dicarboxylic acids
Alpha-N-phenylacetyl-L-glutamine	<0.001	Amino acids	2,5-Furandicarboxylic acid	<0.001	
Indolelactic acid	<0.001	Indolyl carboxylic acids	5-Hydroxyindoleacetic acid	<0.001	Indole-3-acetic acid derivatives
Isobutyric acid	<0.001	Branched FA	PC(16:0/16:0)	<0.001	PC
N-Methylhydantoin	<0.001	Hydantoins	1,3-Dimethyluric acid	<0.001	Xanthines
Homoveratric acid	<0.001	Phenylacetic acids	Paracetamol sulfate	<0.001	Phenylsulfates
Acetaminophen	<0.001	Aminophenols	Cis,cis-Muconic acid	<0.001	
Scopolamine	<0.001		Citrulline	<0.001	Amino acids
L-Kynurenine	<0.001	Butyrophenones	L-Acetylcarnitine	<0.001	Acyl carnitines
Caffeine	<0.001	Xanthines	Guanidoacetic acid	<0.001	Amino acids
Deoxycholic acid glycine conjugate	<0.001	C24 bile acids	Acetic acid	<0.001	Saturated FA
3-Hexenedioic acid	<0.001	Dicarboxylic acids	N-Acetylserotonin	<0.001	Serotonins
Salicyluric acid	<0.001	Hippuric acids	Niacinamide	<0.001	Nicotinamides
L-Arabinose	<0.001	Monosaccharides	Indole-3-carboxylic acid	<0.001	Indolecarboxylic acids
Niacinamide	<0.001	Nicotinamides	Indole-3-carboxylic acid	<0.001	Indolecarboxylic acids
3,4,5-Trimethoxycinnamic acid	<0.001	Coumaric acids	Uridine	0.009	Pyrimidine ribonucleosides
Adenine	<0.001	6-aminopurines	L-Glutamine	0.001	Amino acids
Acetaminophen glucuronide	<0.001	Sugar acids	Acetone	0.001	Ketones
3-Methoxyphenylacetic acid	<0.001		Naproxen	0.002	Naphthalenes
Oxypurinol	<0.001	Xanthines	DUMP	0.002	Pyrimidine dNMP

Metabolite	FDR	Subclass	Metabolite	FDR	Subclass
Kynurenic acid	<0.001	Quinoline carboxylic acids	Phenol	0.003	Phenols
Hypoxanthine	<0.001	Hypoxanthines	Uric acid	0.005	Xanthines
1-Methyladenosine	<0.001	Purine ribonucleosides	5-Aminolevulinic acid	0.007	Amino FA
Nicotinamide ribotide	<0.001	Nicotinamide nucleotides	Gallic acid	0.009	Gallic acids
Nicotinuric acid	<0.001	Amino acids			

Table 5-4: List of significant metabolites between uncontrolled diabetics and prediabetics/controlled diabetics (Wilcoxon rank-sum test).

Metabolite	FDR	Subclass	Metabolite	FDR	Subclass
3-Indolepropionic acid	<0.001	Indolyl carboxylic acids	**Glucosamine**	0.006	Amino sugars
2-Phenylbutyric acid	<0.001		**Thyroxine**	0.008	Diarylethers
Hexadecanedioic acid	<0.001	Dicarboxylic acids	**L-Arabinose**	0.014	Monosaccharides
5-Dodecenoic acid	<0.001	Unsaturated FA	**3-Methylxanthine**	0.019	Xanthines
Homovanillic acid	<0.001	Phenylacetic acids	**Homocysteine**	0.019	Amino acids
Succinylacetone	<0.001	Medium-chain keto acids	**ADP**	0.020	Purine rNDP
Umbelliferone	<0.001	Hydroxycoumarins	**Creatine**	0.023	Amino acids
Indoleacetic acid	<0.001	Indole-3-acetic acid derivatives	**L-Valine**	0.023	Amino acids
Cortisol	<0.001	C21 steroids	**Guanine**	0.025	Hypoxanthines
Dodecanedioic acid	<0.001	Dicarboxylic acids	**Isobutyric acid**	<0.001	Branched FA
Indolelactic acid	<0.001	Indolyl carboxylic acids	**3,5-Dimethoxyphenol**	<0.001	
Homoveratric acid	<0.001	Phenylacetic acids	**L-Acetylcarnitine**	0.001	Acyl carnitines
4-Ethylbenzoic acid	<0.001	Benzoic acids	**Naproxen**	0.004	Naphthalenes
Glycocholic acid	<0.001	C24 bile acids	**Acetic acid**	0.004	Saturated FA
L-Kynurenine	<0.001	Butyrophenones	**Glycylproline**	0.005	Dipeptides
Nutriacholic acid	<0.001		**Glucosamine**	0.006	Amino sugars
PC(16:0/16:0)	<0.001	PC	**Thyroxine**	0.008	Diarylethers
3-Methoxyphenylacetic acid	<0.001		**L-Arabinose**	0.014	Monosaccharides
Cis,cis-Muconic acid	<0.001		**3-Methylxanthine**	0.019	Xanthines
PC(18:1(9Z)/18:1(9Z))	<0.001	PC	**Homocysteine**	0.019	Amino acids
Deoxycholic acid glycine conjugate	<0.001	C24 bile acids	**ADP**	0.020	Purine rNDP
Traumatic acid	<0.001	Dicarboxylic acids	**Creatine**	0.023	Amino acids
5-Hydroxyindoleacetic acid	<0.001	Indole-3-acetic acid derivatives	**L-Valine**	0.023	Amino acids
Aniline	<0.001	Anilines	**Guanine**	0.025	Hypoxanthines

Metabolite	FDR	Subclass	Metabolite	FDR	Subclass
Inosinic acid	<0.001	Purine rNMP	**5-Methoxytryptophol**	<0.001	Indoles
1-Butanol	<0.001	Fatty alcohols	**Adenine**	<0.001	6-aminopurines
Elaidic acid	<0.001	Unsaturated FA	**Aspartame**	<0.001	Peptides
3,4,5-Trimethoxycinnamic acid	<0.001	Coumaric acids	**Pantothenic acid**	<0.001	Amino acids
Adenosine monophosphate	<0.001	Purine rNMP	**L-Norleucine**	<0.001	Amino FA
1-Methyladenosine	<0.001	Purine ribonucleosides			

5.3.5 Analysis of metabolic pathway

Since the metabolomics assay captured a high dimensionality snapshot of the state of the respective metabolome, as shown in Tables 5-2, 5-3, and 5-4, the focus was on the over-represented subsets in the outcomes. PA is to reduce dimensionality and ease functional interpretation. Based on existing knowledge of biological pathways, molecular entities such as metabolites can be mapped onto curated pathway sets to represent how these entities collectively function and interact in a biological context [109]. The MetaboAnalyst 5.0 platform performed pathway enrichment and topological analysis of differential metabolites in blood. Groupwise PA was performed assuming the three groups indicated in Figure 5-4 B (non-diabetics, prediabetics, and controlled diabetics, and uncontrolled diabetics). Based on the identified metabolites in our data, the MetaboAnalyst 5.0 platform detected 39 metabolic pathways for each groupwise comparison as mentioned above. These metabolic pathways are shown in Figures 5-8 and 5-9. In Figure 5-8, total compounds are the total number of compounds in the KEGG library pathway; the Hits are the actually matched number from our uploaded data. In Figure 5-9, potential target pathways were screened according to each pairwise comparison's log(P) value and pathway impact score. The metabolome view shows all matched pathways according to the p values from the pathway enrichment analysis and pathway impact values from the pathway topology analysis. Each bubble in the bubble diagram represents a metabolic pathway. Color gradient and circle size indicate the significance of the pathway ranked by P-value (yellow: higher P-values and red: lower P-values) and pathway impact score (the larger the circle, the higher the pathway impact score). According to the -log(P) value and pathway impact score, the top metabolic pathways were identified by name.

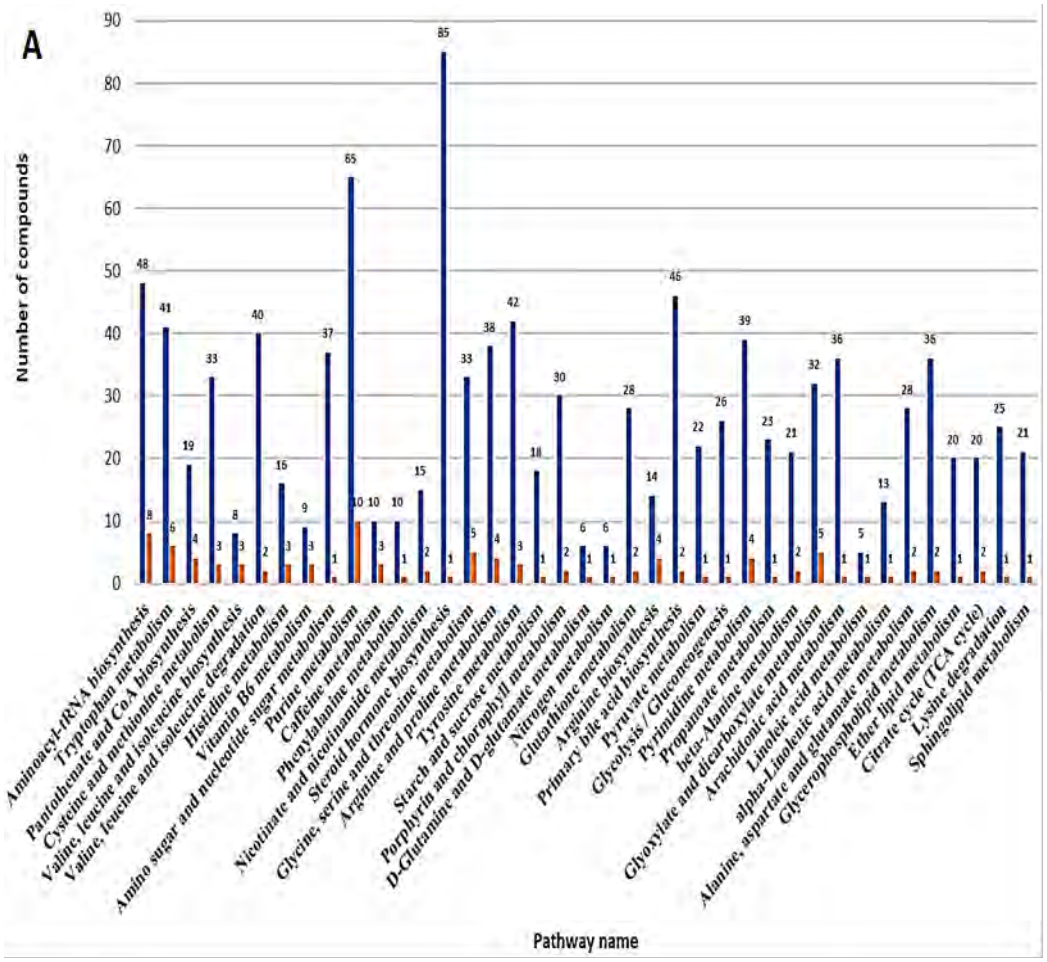
The arbitrary FDR cut-off values (fdrcut-off) are 0.00007, 0.000038, and 0.0417 for non-diabetics vs. uncontrolled diabetics, non-diabetics vs. pre-diabetics and controlled diabetics, and uncontrolled diabetics vs. pre-diabetics and controlled diabetics, respectively. However, there was no considerable increase in significantly enriched pathways when the fdrcut-off was increased beyond the preferred values. For non-diabetics vs. uncontrolled diabetics, the top fifteen metabolic pathways were selected as shown in Figure 5-9 A: Aminoacyl-tRNA biosynthesis, Tryptophan metabolism, Pantothenate, and CoA biosynthesis, Cysteine and methionine metabolism, Valine, leucine and isoleucine dégradation, Valine, leucine and isoleucine biosynthèses, Histidine metabolism, Vitamin B6 metabolism, Caffeine metabolism, Purine metabolism, Amino sugar and nucleotide sugar metabolism, Phenylalanine metabolism, Nicotinate and nicotinamide metabolism, Steroid hormone biosynthesis and Glycine, serine and threonine metabolism. Metabolic PA indicated 42 different metabolites enriched in these fifteen metabolic pathways. Wilcoxon rank-sum test analyzed the differential metabolites enriched in the identified pathways. The levels of 5-Hydroxyindoleacetic acid ($P < 0.001$), Paraxanthine ($P < 0.001$), L-Tryptophan ($P < 0.001$), Deoxyguanosine ($P < 0.001$), L-Kynurenine ($P < 0.001$), Pantothenic acid ($P < 0.001$), Guanine ($P < 0.001$), Alpha-ketoisovaleric acid ($P < 0.001$), Urocanic acid ($P < 0.001$), Pyridoxal ($P < 0.001$), N-acetylserotonin ($P < 0.001$), Homocysteine ($P < 0.001$), Indoleacetic acid ($P < 0.001$), 5-Hydroxy-L-tryptophan ($P < 0.001$), Glucosamine ($P < 0.001$), Guanidoacetic acid ($P < 0.001$), Cortisol ($P < 0.001$), Hippuric acid ($P < 0.001$), Caffeine ($P < 0.001$), 4-Pyridoxic acid ($P < 0.001$), Pyridoxal 5'-phosphate ($P < 0.001$), Urea ($P < 0.001$), Nicotinamide ribotide ($P < 0.001$), Adenine ($P < 0.001$), ADP ($P < 0.001$), Hypoxanthine ($P < 0.001$), Adenosine monophosphate ($P < 0.001$), L-Glutamine ($P < 0.001$), 5-Aminolevulinic acid ($P < 0.001$), 1-Methylhistidine ($P < 0.001$), Glycine ($P < 0.001$), and Uric acid ($P = 0.026$) were significantly different among non-diabetics and uncontrolled diabetics. Boxplots of the identified significant metabolites for non-diabetics vs. uncontrolled diabetics groups are shown in Figure 5-10. However, L-Histidine, levels of L-Arginine, L-methionine, L-Valine, L-Proline, Uracil, 2-Ketobutyric acid, Theobromine, Inosinic acid and Creatine are insignificantly different among ND and Uncontrolled D.

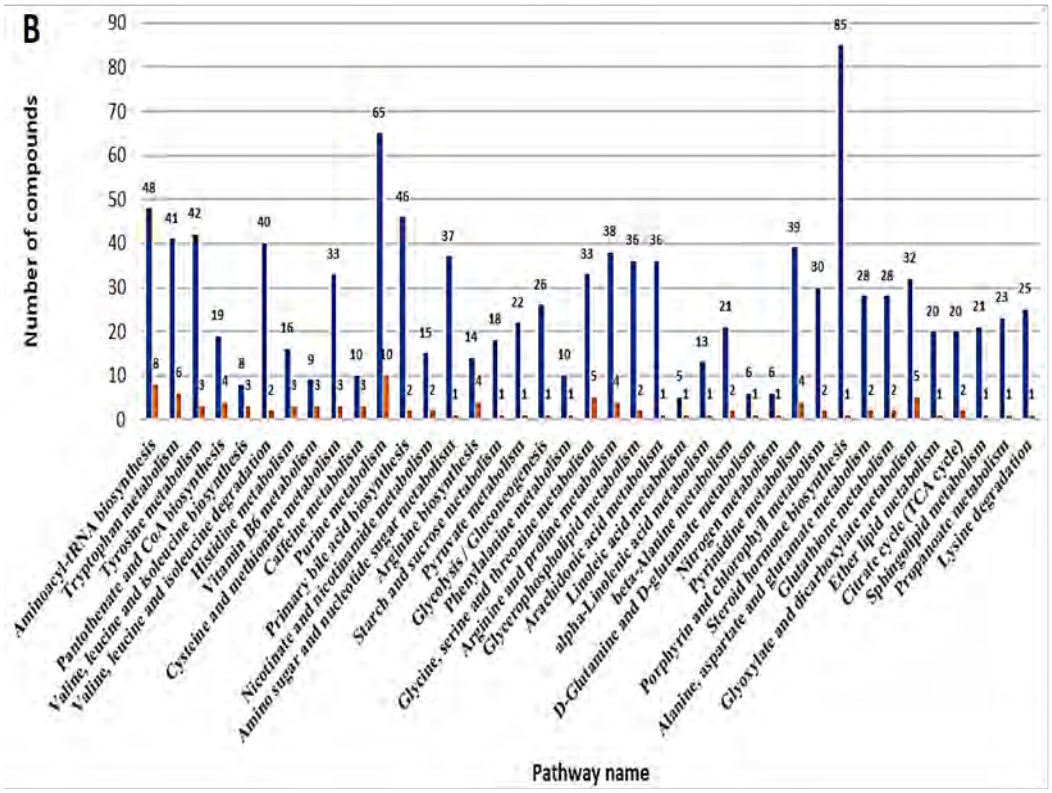
The same analysis was repeated for non-diabetics vs. pre-diabetics and controlled diabetics, the top twenty one metabolic pathways were selected as shown in Figure 5-9

B: Aminoacyl-tRNA biosynthesis, Tryptophan metabolism, Pantothenate and CoA biosynthesis, Tyrosine metabolism, Cysteine and methionine metabolism, Valine, leucine and isoleucine dégradation, Valine, leucine and isoleucine biosynthèses, Histidine metabolism, Vitamin B6 metabolism, Caffeine metabolism, Purine metabolism, Amino sugar and nucleotide sugar metabolism, Phenylalanine metabolism, Primary bile acid biosynthesis, Nicotinate and nicotinamide metabolism, Starch and sucrose metabolism, Arginine biosynthesis, Pyruvate metabolism, Glycolysis / Gluconeogenesis, Glycine, serine and threonine metabolism and Arginine and proline metabolism . There were 48 enriched metabolites in the selected pathways. The levels of of Paraxanthine ($P < 0.001$), Hippuric acid ($P < 0.001$), Glucosamine ($P < 0.001$), Glycocholic acid ($P < 0.001$), Trehalose ($P < 0.001$), Pantothenic acid ($P < 0.001$), Guanidoacetic acid ($P < 0.001$), N-Acetylserotonin ($P < 0.001$), 5-Hydroxyindoleacetic acid ($P < 0.001$), 5-Hydroxy-L-tryptophan ($P < 0.001$), L-Kynurenine ($P < 0.001$), epinephrine ($P < 0.001$, Homovanillic acid ($P < 0.001$), Citrulline ($P < 0.001$), Niacinamide ($P < 0.001$), Thyroxine ($P < 0.001$), L-Tryptophan ($P < 0.001$), alpha-Ketoisovaleric acid ($P < 0.001$), Urocanic acid ($P <0.001$), 1-Methylhistidine ($P < 0.001$), Pyridoxal 5'-phosphate ($P < 0.001$), Pyridoxal ($P < 0.001$), 4-Pyridoxic acid ($P < 0.001$), Homocysteine ($P < 0.001$), Paraxanthine ($P < 0.001$), Caffeine ($P < 0.001$), L-Glutamine ($P = 0.002$), Nicotinamide ribotide ($P < 0.001$), ADP ($P = 0.023$), Adenosine monophosphate ($P < 0.001$), Inosinic acid ($P = 0.012$) , Hypoxanthine ($P < 0.001$), 5-Aminolevulinic acid ($P < 0.001$),Guanine ($P < 0.001$), Deoxyguanosine ($P < 0.001$), Adenine ($P = 0.001$), Uric acid ($P = 0.003$), Urea ($P< 0.001$), Creatine ($P = 0.020$), Glycine($P = 0.030$) and Inosinic acid ($P = 0.008$) are significant among non-diabetics vs. pre-diabetics and controlled diabetics. Boxplots of the identified significant metabolites for non-diabetics vs. pre-diabetics and controlled diabetics groups are presented in Figure 5-11. On the other hand, levels of L-Histidine, L-Arginine, L-Proline, L-Valine, L-Methionine, 2-Ketobutyric acid, Uracil, and Theobromin are the same for non-diabetics vs. pre-diabetics and controlled diabetics.

Lastly, a groupwise comparison between uncontrolled diabetics vs. prediabetics and controlled diabetics was conducted. The top twelve metabolic pathways were selected as shown in Figure 5-9 C: Tyrosine metabolism, Tryptophan metabolism, Primary bile acid biosynthesis, Steroid hormone biosynthesis, Glycerophospholipid metabolism, Arachidonic acid metabolism, Linoleic acid metabolism, Pyruvate metabolism, Purine

metabolism, Glycolysis / Gluconeogenesis, alpha-Linolenic acid metabolism, and Vitamin B6 metabolism. There were 29 enriched metabolites in the identified pathways. The level of Homovanillic acid ($P < 0.001$), Indoleacetic acid ($P < 0.001$), Cortisol ($P < 0.001$), Glycocholic acid ($P < 0.001$), L-Kynurenine ($P < 0.001$), PC(16:0/16:0)($P < 0.001$), PC(18:1(9Z)/18:1(9Z))($P < 0.001$), Inosinic acid ($P < 0.001$), 5-Hydroxyindoleacetic acid ($P < 0.001$), Adenosine monophosphate ($P < 0.001$), Adenine ($P < 0.001$), Acetic acid ($P < 0.001$), Thyroxine ($P < 0.001$), ADP ($P = 0.010$), Guanine ($P = 0.018$) are significantly different between uncontrolled diabetics vs. pre-diabetics and controlled diabetics. Boxplots of the identified significant metabolites for uncontrolled diabetics vs. prediabetics and controlled diabetics groups are presented in Figure 5-12. However, levels of N-Acetylserotonin, Deoxyguanosine, 4-Pyridoxic acid, Pyridoxal 5'-phosphate, Epinephrine, L-Tryptophan, Hypoxanthine, L-Glutamine, Uric acid, Pyridoxal, 5-Hydroxy-L-tryptophan, Glycine, Glycerophosphocholine, and Urea are the same for uncontrolled diabetics vs. prediabetics and controlled diabetics.


A
Number of compounds
Pathway name
Aminoacyl-tRNA biosynthesis
Tryptophan metabolism
Pantothenate and CoA biosynthesis
Cysteine and methionine metabolism
Valine, leucine and isoleucine biosynthesis
Valine, leucine and isoleucine degradation
Histidine metabolism
Vitamin B6 metabolism
Amino sugar and nucleotide sugar metabolism
Purine metabolism
Caffeine metabolism
Phenylalanine metabolism
Nicotinate and nicotinamide metabolism
Steroid hormone biosynthesis
Glycine, serine and threonine metabolism
Arginine and proline metabolism
Tyrosine metabolism
Starch and sucrose metabolism
Porphyrin and chlorophyll metabolism
D-Glutamine and D-glutamate metabolism
Nitrogen metabolism
Glutathione metabolism
Arginine biosynthesis
Primary bile acid biosynthesis
Pyruvate metabolism
Glycolysis / Gluconeogenesis
Pyrimidine metabolism
Propanoate metabolism
beta-Alanine metabolism
Glyoxylate and dicarboxylate metabolism
Arachidonic acid metabolism
Linoleic acid metabolism
alpha-Linolenic acid metabolism
Alanine, aspartate and glutamate metabolism
Glycerophospholipid metabolism
Ether lipid metabolism
Citrate cycle (TCA cycle)
Lysine degradation
Sphingolipid metabolism


B
Number of compounds
Pathway name
Aminoacyl-tRNA biosynthesis
Tryptophan metabolism
Tyrosine metabolism
Pantothenate and CoA biosynthesis
Valine, leucine and isoleucine biosynthesis
Valine, leucine and isoleucine degradation
Histidine metabolism
Cysteine and methionine metabolism
Vitamin B6 metabolism
Caffeine metabolism
Purine metabolism
Primary bile acid biosynthesis
Nicotinate and nicotinamide metabolism
Amino sugar and nucleotide sugar metabolism
Arginine biosynthesis
Starch and sucrose metabolism
Pyruvate metabolism
Glycolysis / Gluconeogenesis
Phenylalanine metabolism
Glycine, serine and threonine metabolism
Arginine and proline metabolism
Glycerophospholipid metabolism
Arachidonic acid metabolism
Linoleic acid metabolism
alpha-Linolenic acid metabolism
D-Glutamine and D-glutamate metabolism
beta-Alanine metabolism
Nitrogen metabolism
Pyrimidine metabolism
Porphyrin and chlorophyll metabolism
Steroid hormone biosynthesis
Alanine, aspartate and glutamate metabolism
Glutathione metabolism
Glyoxylate and dicarboxylate metabolism
Ether lipid metabolism
Citrate cycle (TCA cycle)
Sphingolipid metabolism
Propanoate metabolism
Lysine degradation

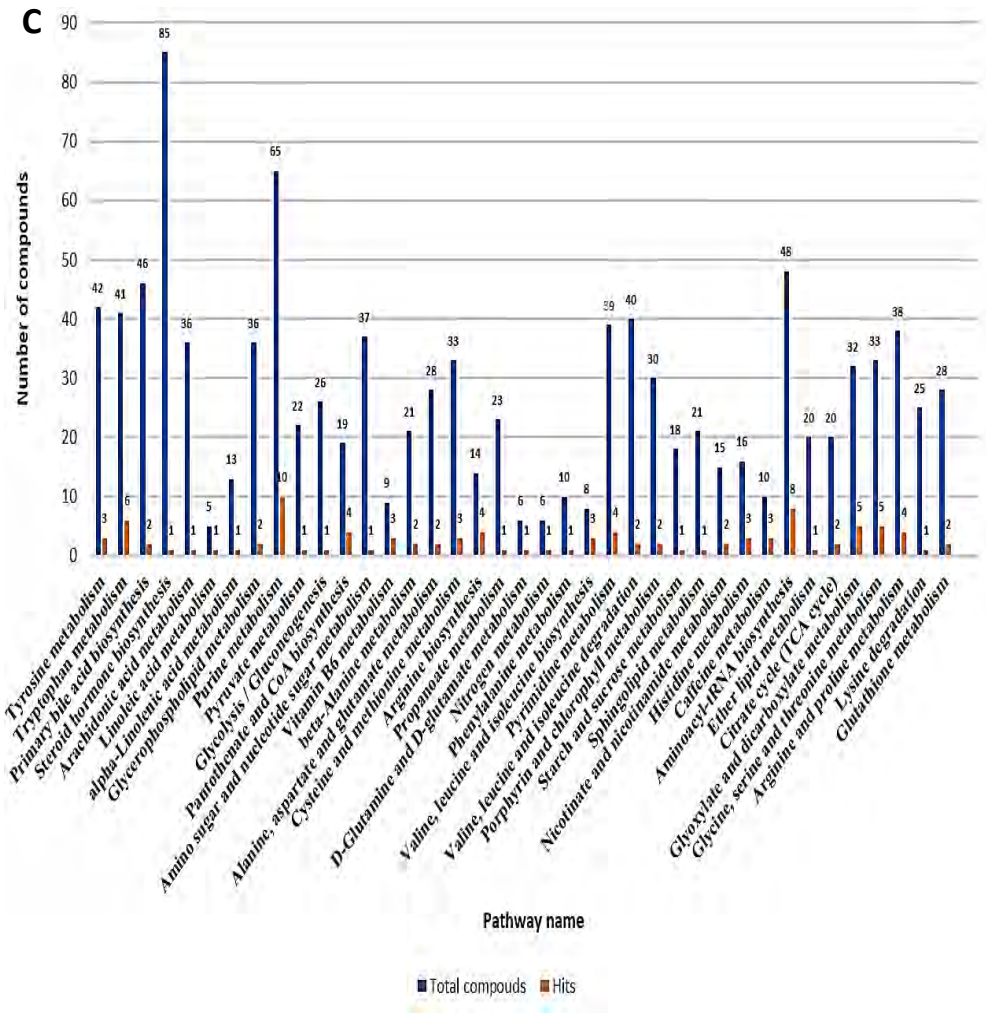


Figure 5-8: Pathway analysis results showing total compounds in each pathway versus the number of matched metabolites from our datasets. (A) Metabolic pathway analysis for ND and Uncontrolled D. (B) Metabolic pathway analysis for ND and Pre/controlled D. (C) Metabolic.

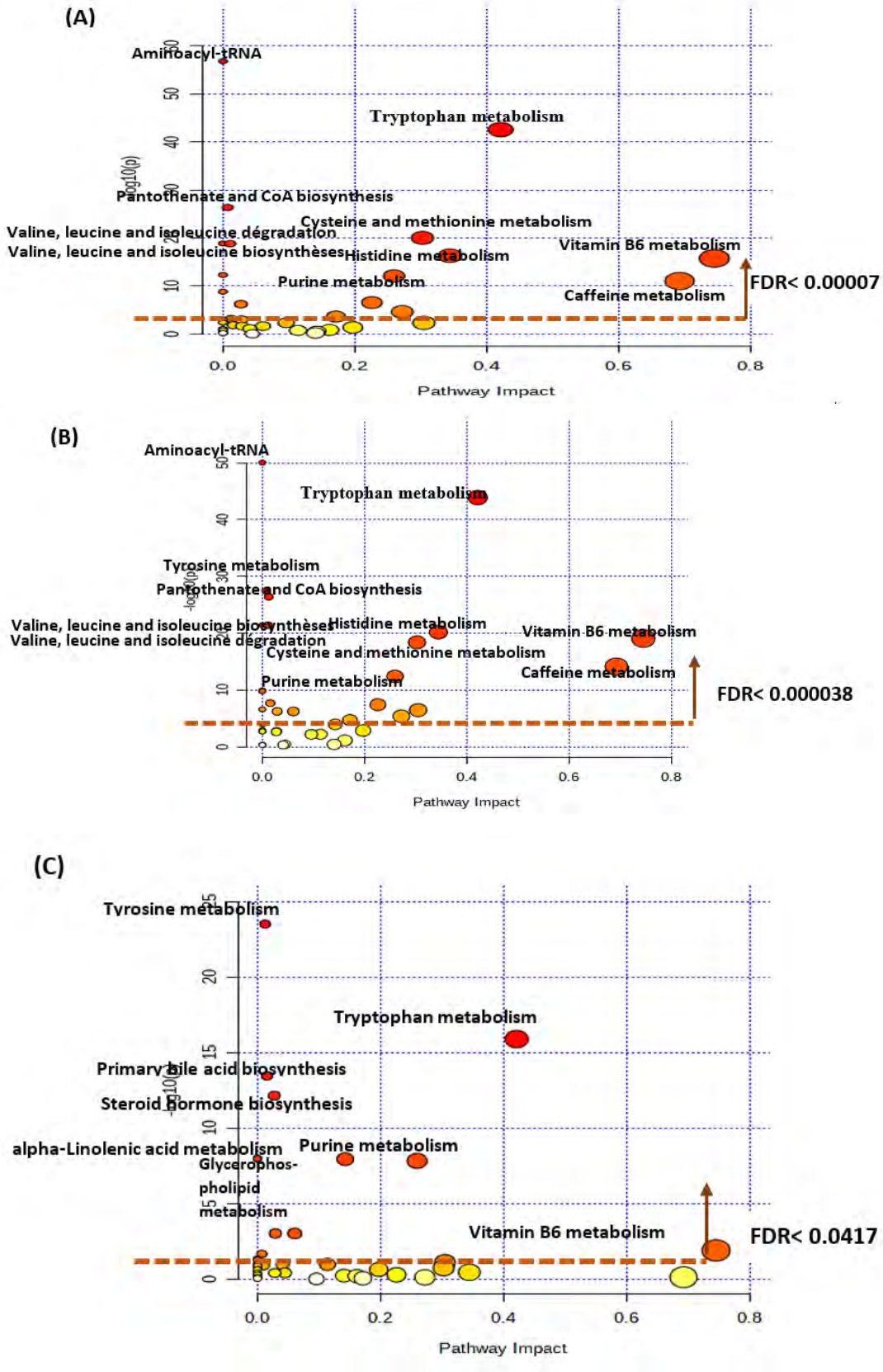


Figure 5-9: Overview of metabolic pathway analysis. (A) Metabolic pathway analysis for ND and Uncontrolled D. (B) Metabolic pathway analysis for ND and Pre/controlled D. (C) Metabolic pathway analysis for Uncontrolled D and Pre/controlled D.

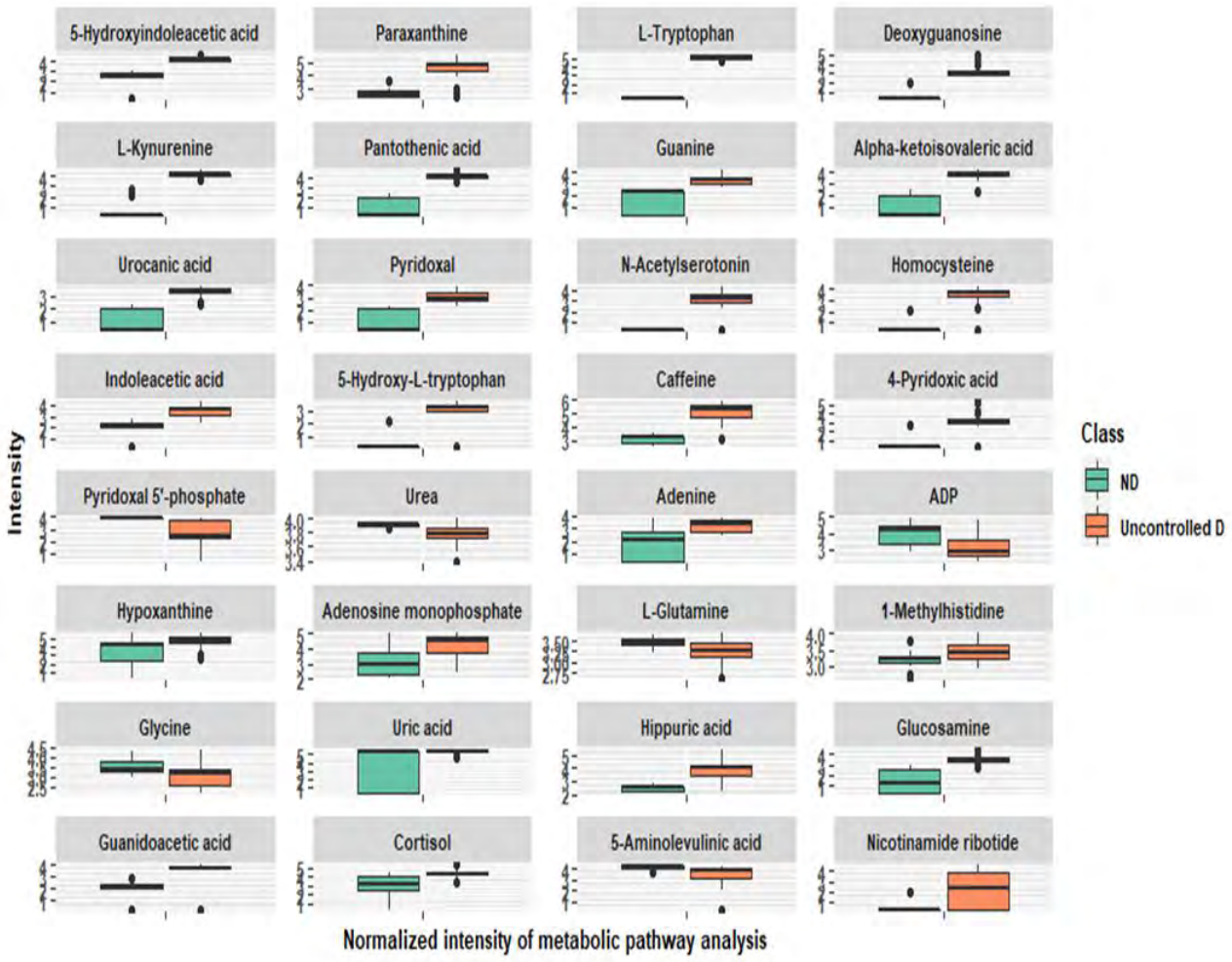


Figure 5-10: Boxplot of normalized intensity metabolites for ND and Uncontrolled D.

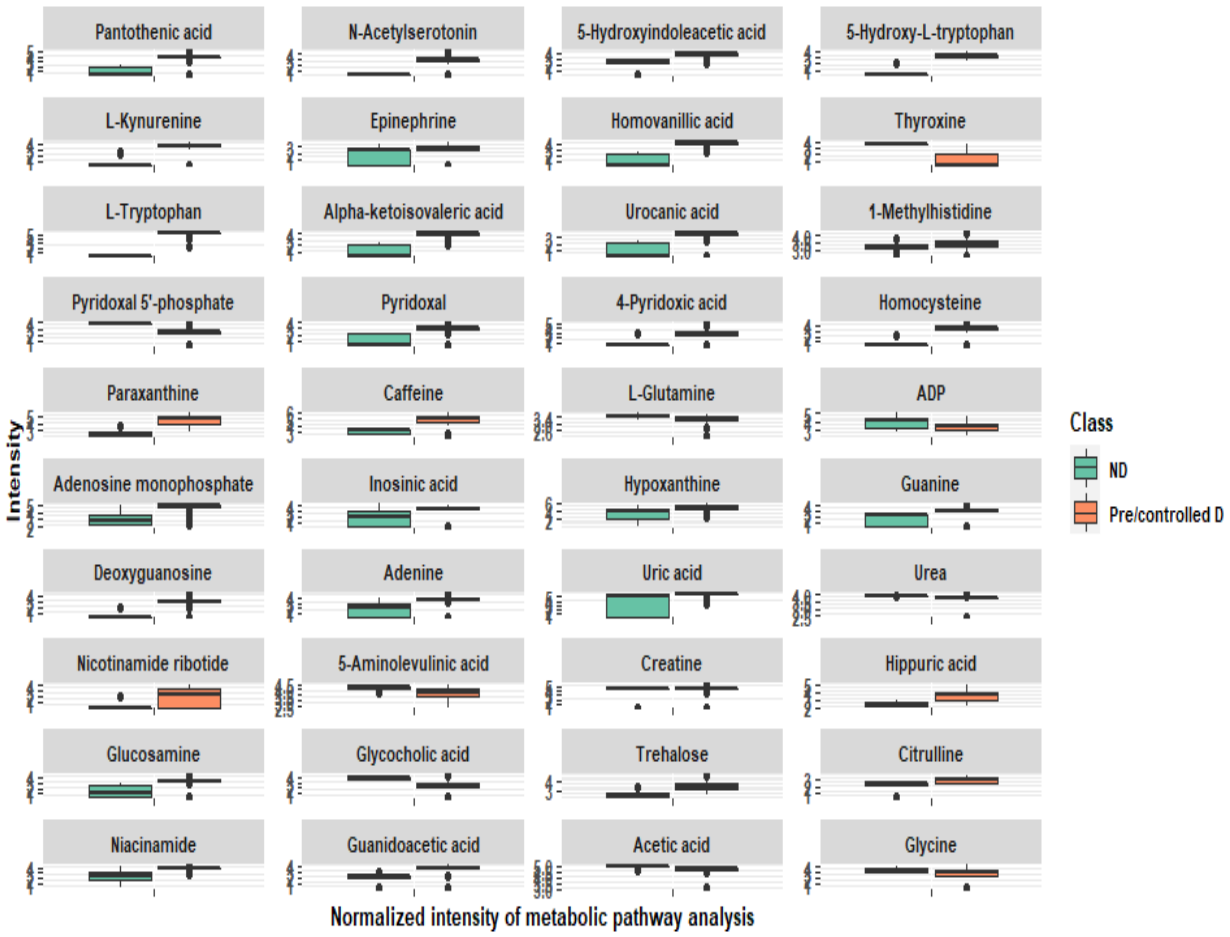


Figure 5-11: Boxplot of normalized intensity metabolites for ND and Pre/controlled D.

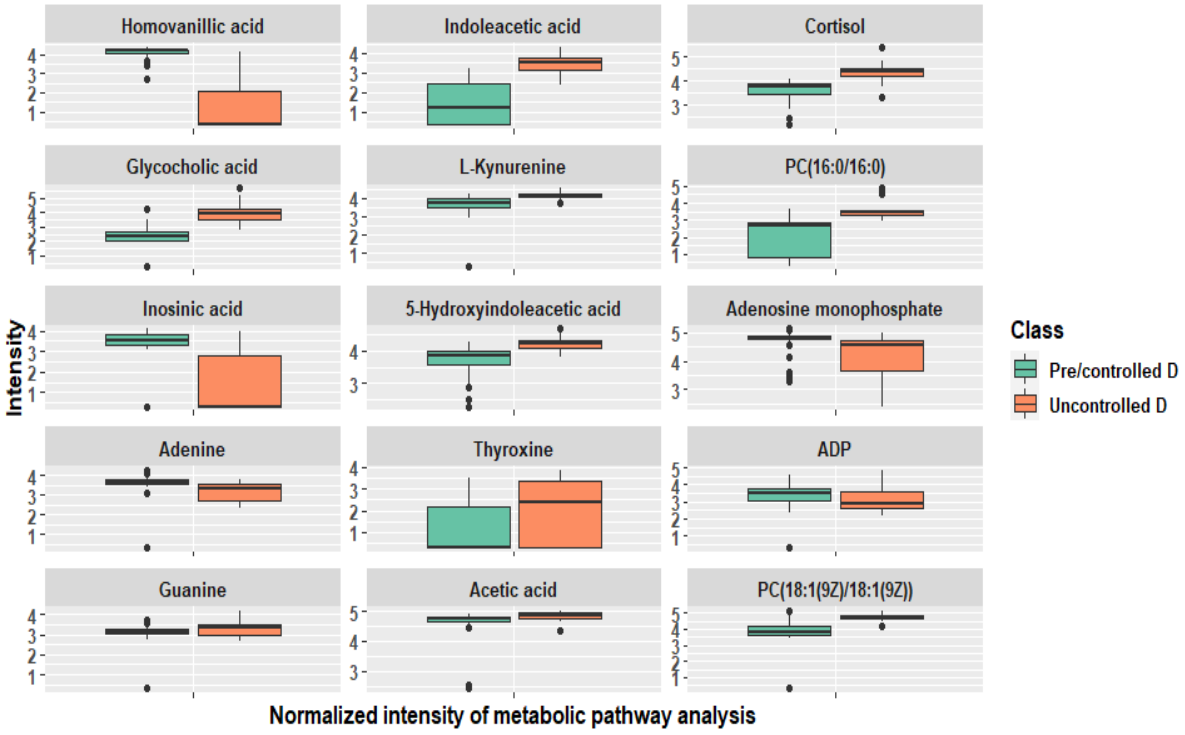


Figure 5-12: Boxplot of normalized intensity metabolites for Uncontrolled D and Pre/controlled D.

5.4 Discussion and Conclusions

This research used LC-MS/MS to analyze blood metabolites of T2DM and non-T2DM Emirati patients. The state-of-the-art LC-MS/MS is a powerful quantitative technique of biological compounds with high specificity, sensitivity, and throughput [245].

The descriptions of dysglycemia, when applied to a continuous pathophysiologic process, may unintentionally undervalue individuals at risk for disease progression. Even within the normal range, the gradual rise of glucose levels happens relatively late in the T2DM advance when β-cell function might already be decreased [246]. Therefore, it is significant to discover biomarkers that predict evolution to dysglycemic states at the earliest point when the β-cell function is still relatively optimal and may be more responsive to lifestyle change [65].

This study has observed fifteen significant potential biomarkers between uncontrolled diabetics against prediabetics and controlled diabetics. The metabolites-associated alterations in fatty acid-, Purines-, bile acid-, Steroids-, Glycerophosphocholines-, Diarylethers-, Phenylacetic acids-, and Indoles-metabolism are identified. Uncontrolled diabetics patients have cortisol excess compared to prediabetics and controlled diabetics vs. non-diabetics. The presence of T2DM and other diseases as chronic complications of cortisol excess, both overt and subclinical, has been established for several years [247-249]. As a result, there is importance in manipulating glucocorticoid action as a therapeutic strategy [247]. Also, a study [250] found that alterations in diurnal cortisol patterns were predictive of future glucose disturbance.

Bile acids (BAs) have appeared as vital signaling molecules in glucose metabolic management [251, 252]. Alterations in BA metabolism have been partially proved in T2DM patients [253]. We also recognized higher levels of Glycocholic acid in uncontrolled diabetics than prediabetics and controlled diabetics. Bile acids might regulate glucose tolerance, insulin sensitivity, and energy metabolism, meaning bile acids may represent a potential therapeutic target for T2DM [254, 255].

Tryptophan, an aromatic amino acid metabolite, has expansive physiological functions regulating growth and feed intake, mood and behavior, and immune responses [256]. Moreover, it is one of the essential amino acids for humans. The ability of tryptophan to identify high-risk individuals before the onset of T2DM and even before significant variations of metabolic markers is noteworthy [257]. Our analysis indicated that

tryptophan metabolites such as 5-Hydroxyindoleacetic acid is higher in diabetic and potentially diabetic patients. Previous studies [258, 259] found that levels of 5-Hydroxyindoleacetic acid were increased in plasma from DM patients, indicating that the metabolism of tryptophan is accelerated in DM patients and that tryptophan metabolism may be altered in T2DM patients due to the many stresses to which T2DM patients are exposed.

Purines are fundamental parts of nucleotides and nucleic acids, playing several essential roles in human physiology, disturbing tissue function, cell integrity, and oxidation. Purine metabolism comprises the synthesis and degradation of purine nucleotides and regulates the adenylate and guanylate pool [260]. Previous studies presented a potential functional connection between Purines metabolites and the onset of T2DM [261, 262]. Our research found Adenine to be higher in prediabetics and controlled diabetics than uncontrolled diabetics and non-diabetics. However, Guanine is higher in the uncontrolled diabetics and prediabetics and controlled diabetics groups than non-diabetics. Also, the non-diabetics group has higher ADP levels than the uncontrolled diabetics and prediabetics and controlled diabetics groups.

On the contrary, the non-diabetics group has lower Caffeine levels than the uncontrolled diabetics and prediabetics and controlled diabetics groups. We also noticed lower Hypoxanthine levels in the non-diabetics group versus the prediabetics and controlled diabetics group. Moreover, Deoxyguanosine shows low levels in non-diabetics versus uncontrolled diabetics and prediabetics and controlled diabetics groups.

Thyroid dysfunction and DM are closely related [263]. In addition, several studies have recognized the increased incidence of thyroid disorders in DM and vice versa [264]. Therefore, thyroid dysfunction is more prevalent in T2DM patients than in general. Consequently, it is recommended that insulin treatment should be adjusted in patients with diabetes after the occurrence of thyroid dysfunction [263]. Our analysis shows that Thyroxine in prediabetics and controlled diabetics has lower values than both the uncontrolled diabetics and non-diabetics groups. Interestingly, a study found that lower thyroid function is a risk factor for incident diabetes, especially in prediabetes [265].

Short-chain fatty acids (SCFAs) were mainly identified in reducing serum glucose levels, improving insulin resistance, mitigating inflammation [266-269]. In addition,

we found Acetic acid with higher levels in uncontrolled diabetics than prediabetics and controlled diabetics.

Amino acids are essential modulators of glucose metabolism, insulin secretion, and insulin sensitivity [270, 271]. A peptide is a short chain of amino acids. The amino acids in a peptide are connected in a sequence by bonds called peptide bonds. Our study recognized that Pantothenic acid has lower non-diabetics values than uncontrolled diabetics and prediabetics and controlled diabetics, consistent with a previous study [272]. On the other hand, Creatine, and Citrulline, changed between non-diabetics and prediabetics, and controlled diabetics. However, amino acids were identified by a previous T2DM UAE study [149] being the most significant metabolites.

In contrast, Glycine is higher in non-diabetics than in prediabetics, controlled diabetics, and uncontrolled diabetics. A previous work [273] stated these findings by reviewing metabolites with altered profiles in individual diabetes. Table 5-5 summarizes diabetes-related metabolites in our study. This study identified differences in metabolites in response to T2DM, agreeing with many published population studies. However, as shown above, the metabolite variation between different studies and our results might have been charged due to specific population constraints such as demographic factors. Therefore, understanding the role of purines and amino acid metabolites in the UAE population needs more investigation as more of our results lie within these categories. Furthermore, linking these metabolites' dysregulation with diabetes pathogenesis is expected to help unlock the triggers of diabetes' high prevalence in the UAE.

Briefly, in the current study, the panel of (43) metabolic signatures can be broadly classified into three pathways: (1) carbohydrate metabolism, (2) amino acid metabolism, and (3) lipid metabolism.

These findings were examined in a single, small cohort, encouraging the need for independent validation in well-designed, largescale studies in the future. Moreover, a targeted metabolomics study can be conducted to validate the discovered metabolites in the study.

Metabolic PA has limitations that might lead to many false-positive pathways. Pathway database choice led to considerably different results in the number and function of significantly enriched pathways [203]. It is also claimed that the selection of pathway databases used in enrichment analyses can have a much stronger effect on the

enrichment results than the statistical corrections used in these analyses [244]. We used the KEGG database library; therefore, it is anticipated to have different effects using other databases such as Reactome and BioCyc.

To conclude, metabolomics is a powerful evolving technology to enhance wellbeing. Combining biomarkers in a clinical setting may provide better sensitivity and specificity in predicting prediabetes and diabetes [65, 66]. Biomarkers offer the ability to identify people with subclinical disease before developing overt clinical disorders [67]. They enable preventive measures to be applied at the subclinical stage and the responses to prophylactic or therapeutic measures to be monitored.

Biomarkers have various applications, including disease detection, diagnosis, prognosis, prediction of response to intervention, and disease monitoring.

Biomarkers could help in daily practice as a diagnostic tool, monitor therapy response, assess prognosis, and as an early marker of disease damage or stratify risk. Biomarkers can also predict drug efficacy more quickly than conventional clinical endpoints. Potential to accelerate product development in specific disease areas. Compared to the more traditional drug-discovery approach, biomarker-enabled drug discovery promotes a better understanding of the disease during target discovery. Biomarkers allow the measurement of drug activity and safety using an endpoint integrated into the drug's therapeutic action.

Table 5-5: Diabetes-related significant metabolites in our study.
↑: indicates that the metabolite is upregulated (increased) from group A-B, ↓: indicates that the metabolite is downregulated (decreased) from group A-B.

Metabolite	**Class of Compound**	**Nature of Variation (Non-Diabetic - Prediabetic)**	**Nature of Variation (Non-Diabetic - Diabetics)**	**Nature of Variation (Prediabetics - Diabetics)**	**Associated Pathway**
Cortisol	Steroids (Sterol Lipids)	-	↑	↑	Steroid hormone biosynthesis.
Glycocholic acid	Bile acids (Sterol Lipids)	↓	-	↑	Primary bile acid biosynthesis.
5-Hydroxyindoleacetic acid	Indoles	↑	↑	↑	Tryptophan metabolism.
Adenine	Purines (Nucleic acids)	↑	↑	↓	Purine metabolism.
Guanine	Purines (Nucleic acids)	↑	↑	↑	Purine metabolism.
ADP	Purines (Nucleic acids)	↓	↓	↓	Purine metabolism.
Caffeine	Purines (Nucleic acids)	↑	↑	-	Caffeine metabolism.
Hypoxanthine	Purines (Nucleic acids)	↑	↑	-	Purine metabolism.
Deoxyguanosine	Purines (Nucleic acids)	↑	↑	-	Purine metabolism.
Thyroxine	Diarylethers (Benzenoids)	↓	-	↑	Tyrosine metabolism.
Acetic acid	Short Chain Fatty Acids	↓	-	↑	Pyruvate metabolism. Glycolysis / Gluconeogenesis.
Pantothenic acid	Amino acids and peptides	↑	↑	-	Pantothenate and CoA biosynthesis.

Metabolite	**Class of Compound**	**Nature of Variation (Non-Diabetic - Prediabetic)**	**Nature of Variation (Non-Diabetic - Diabetics)**	**Nature of Variation (Prediabetics - Diabetics)**	**Associated Pathway**
Creatine	Amino acids and peptides	↑	-	-	Glycine, serine, and threonine metabolism. Arginine and proline metabolism.
Citrulline	Amino acids and peptides	↑	-	-	Arginine biosynthesis.
Glycine	Amino acids and peptides	↓	↓	-	Primary bile acid biosynthesis. Aminoacyl-tRNA biosynthesis. Glycine, serine and threonine metabolism.
1-Methylhistidine	Amino acids and peptides	↑	↑	-	Histidine metabolism
4-Pyridoxic acid	Pyridines	↑	↑	-	Vitamin B6 metabolism
5-Aminolevulinic acid	Amino acids and peptides	↓	↓	-	Cysteine and methionine metabolism
5-Hydroxy-L-tryptophan	Indoles	↑	↑	-	Tryptophan metabolism
Adenosine monophosphate	Purine nucleotides	↑	↑	↓	Purine metabolism
Alpha-ketoisovaleric acid	Fatty Acids	↑	↑	-	Valine, leucine and isoleucine degradation
Glucosamine	Monosaccharides	↑	↑	-	Amino sugar and nucleotide sugar metabolism
Guanidoacetic acid	Amino acids and peptides	↑	↑	-	Arginine and proline metabolism and Glycine, serine and threonine metabolism
Guanine	Purines	↑	↑	↑	Purine metabolism
Hippuric acid	Benzamides	↑	↑	-	Phenylalanine metabolism
Homocysteine	Amino acids and peptides	↑	↑	-	Cysteine and methionine metabolism
Homovanillic acid	Phenylacetic acids	↑	-	↓	Tyrosine metabolism
Inosinic acid	Purines	↑	-	↓	Purine metabolism
L-Glutamine	Amino acids and peptides	↓	↓	-	Purine metabolism
L-Kynurenine		↑	↑	-	Tryptophan metabolism
N-acetylserotonin	Tryptamines	↑	↑	-	Tryptophan metabolism
Niacinamide	Pyridinecarboxylic acids	↑	-	-	Nicotinate and nicotinamide metabolism
Nicotinamide ribotide	Nicotinamides	↑	↑	-	Nicotinate and nicotinamide metabolism
Paraxanthine	Purines	↑	↑	-	Caffeine Metabolism
PC(16:0/16:0)	Glycerophosphocholines	-	-	↑	Glycerophospholipid metabolism
PC(18:1(9Z)/18:1(9Z))	phosphatidylcholines	-	-	↑	Glycerophospholipid metabolism
Pyridoxal	Pyridine carboxaldehydes	↑	↑	-	Vitamin B6 metabolism
Pyridoxal 5'-phosphate	Pyridine carboxaldehydes	↓	↓	-	Vitamin B6 metabolism
Trehalose	Disaccharides	↑	-	-	Starch and sucrose metabolism
Urea	Carboximidic acids	↓	↓	-	Purine metabolism
Uric acid	purine	↑	↑	-	Purine metabolism
Urocanic acid	Imidazoles	↑	↑	-	Histidine metabolism

Chapter 6. Metabolomic Plasma Profiling of Emirati Dialysis Patients with T2DM versus Non-T2DM

6.1 Introduction

DKD is considered one of the significant complications of DM. However, despite the surge of studies in the field, understanding disease mechanisms is still lacking. Furthermore, the metabolomic profile of DKD under hemodialysis from the middle eastern populations is still unknown. Therefore, this part of our work explores the metabolomic profile of diabetic and non-diabetic UAE citizens undergoing hemodialysis to uncover the potential novel biomarkers in this population.

6.2 Materials and Methods

6.2.1 Patients

We conducted a single-site cross-sectional study, and all available patients were recruited. However, the sample size is constrained by the available resources, such as individuals' willingness to participate and the cost of sample analysis. Therefore, 36 subjects from Emirati citizens who are treated at University Hospital Sharjah were selected.

6.2.2 Sample collection, preparation, and analytical analysis

A total of 4 mL of blood was collected from each patient after overnight fasting into a sterile container. All samples were assembled at roughly the same time each day (between 8 and 10 am every day). The samples preparation method was described previously in the Methods chapter.

TimsTOF Mass Spectrometer (BRUKER, Germany) and MetaboScape software version 4 (Brucker) were employed to separate and detect the cell metabolites. Detailed explanation about LC-MSMS analytical techniques is also found in the Methods chapter.

6.2.3 Statistical and pathway analysis

The statistical approach to analyzing the data is similar to the one used in Chapter 5. First, we found lists of differential metabolites, and then pathway analysis gave a view about the most enriched metabolites and their related- pathways.

6.3 Results

6.3.1 Patients

We enrolled 36 subjects who are being treated at University Hospital Sharjah, UAE. There are 20 females aged between 56 and 85 (average: 69.9 ± 8.16 years; median: 69 years), and 16 males aged between 34 and 90 (average: 73.68 ± 13.07 years; median: 74 years). Out of the 36 participants, 11 are hemodialysis diabetic patients, and 25 are non-diabetic hemodialysis patients. The classification for patients is based on the clinically confirmed diabetic status according to WHO diagnostic criteria for diabetes (fasting plasma glucose ≥ 7.0mmol/l (126mg/dl) or 2–hrs. plasma glucose ≥ 11.1mmol/l (200mg/dl)). However, we will classify the patients based on their most recent HbA1c values for further analysis. The patients are elderly with renal complications of diabetes. Our data show that most known diabetic hemodialysis patients have their diabetes controlled (72.3%). Surprisingly, about 32.0% of the known non-diabetic hemodialysis patients uncontrolled their blood glucose. There is no statistically significant difference in age, gender, HbA1c, cholesterol total, and blood Hb between DD and DND groups ($P > 0.05$).

6.3.2 Differential metabolite screening

Using the LC-MS-MS technique and HMDB database [45], 142 metabolites were identified. These detected and identified metabolites were documented. The top 50 metabolites based on the differences in averages between DD and DND groups are displayed as a heatmap in Figure 6-1. Heatmap in Figure 6-1 shows detected metabolites among DD and DND groups. The color gradient demonstrates concentration levels for each metabolite in each sample. Heatmap in Figure 6-1 indicates no apparent differences in the concentration of the metabolites among the two groups. However, a few of these metabolites show potential differences, such as Alpha-Aspartyl-Lys and Cis-Aconitic Acid.

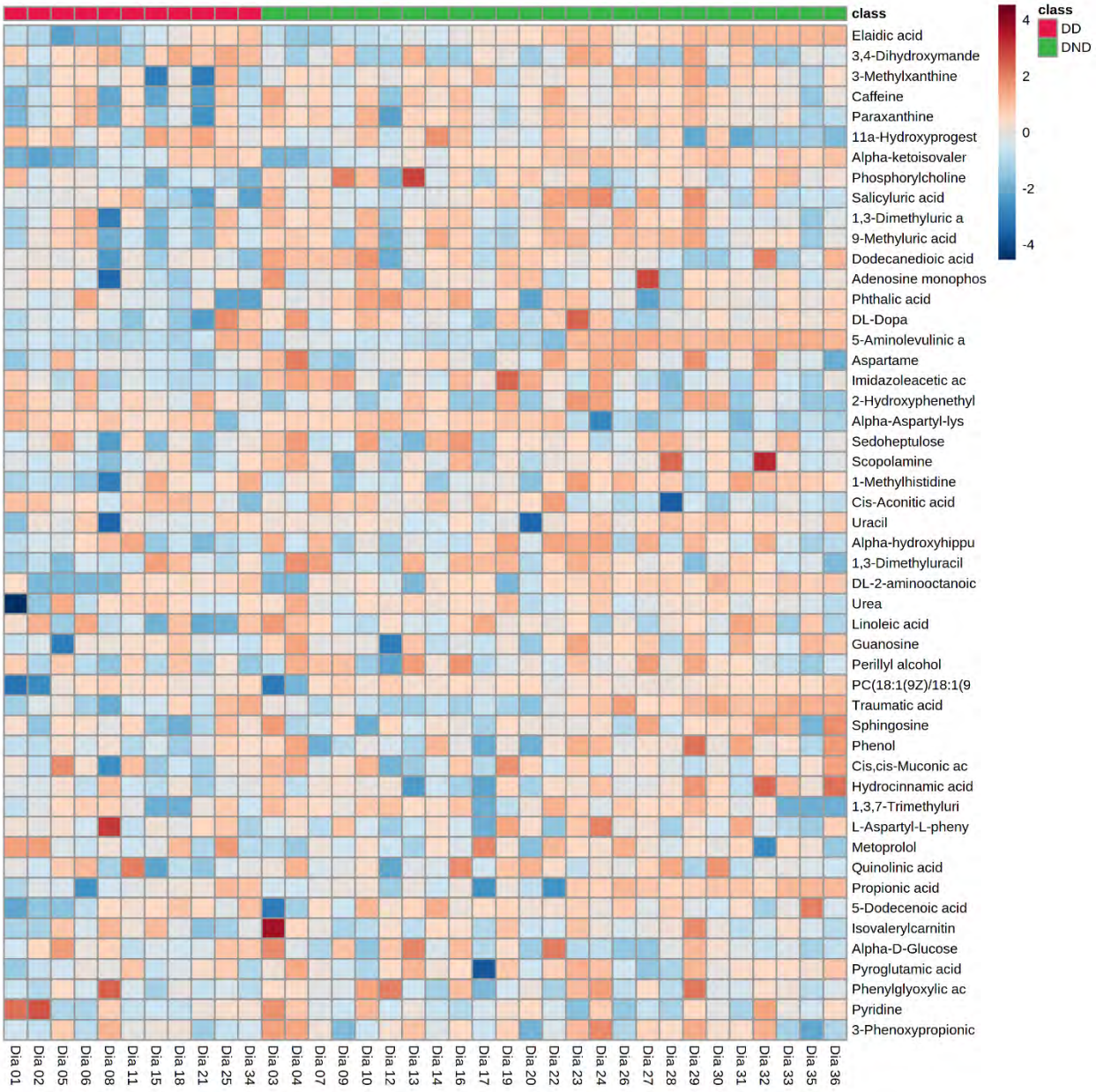


Figure 6-1: Heatmap of the 50 selected metabolites among the DD and DND patients (clinically confirmed diabetic status).

6.3.3 Multivariate statistical analysis

As stated previously, statistical analysis was performed based on (1) clinically confirmed diabetic status (2) HbA1c values. Plots of the top two principal components following the PCA analysis of the 142 identified metabolites under the two scenarios considered are shown in Figure 6-2. Figure 6-2A shows a PCA plot following known diabetic status for patients grouping. The plot in Figure 6-2A depicts that the blood components of the DD group and DND group did not have apparent clustering indicating almost similar metabolic profiles among the two groups. Therefore, both

groups' latest available HbA1c values were used for further PCA analysis. Figure 6-2B shows the PCA plot following participants' grouping based on their latest HbA1c value (controlled if HbA1c value is less than 6.4% and uncontrolled otherwise). Figure 6-2B illustrates an improved separation between controlled and uncontrolled groups. Participants with uncontrolled blood glucose tend to have lower values of PC2 compared to the participants with controlled blood glucose.

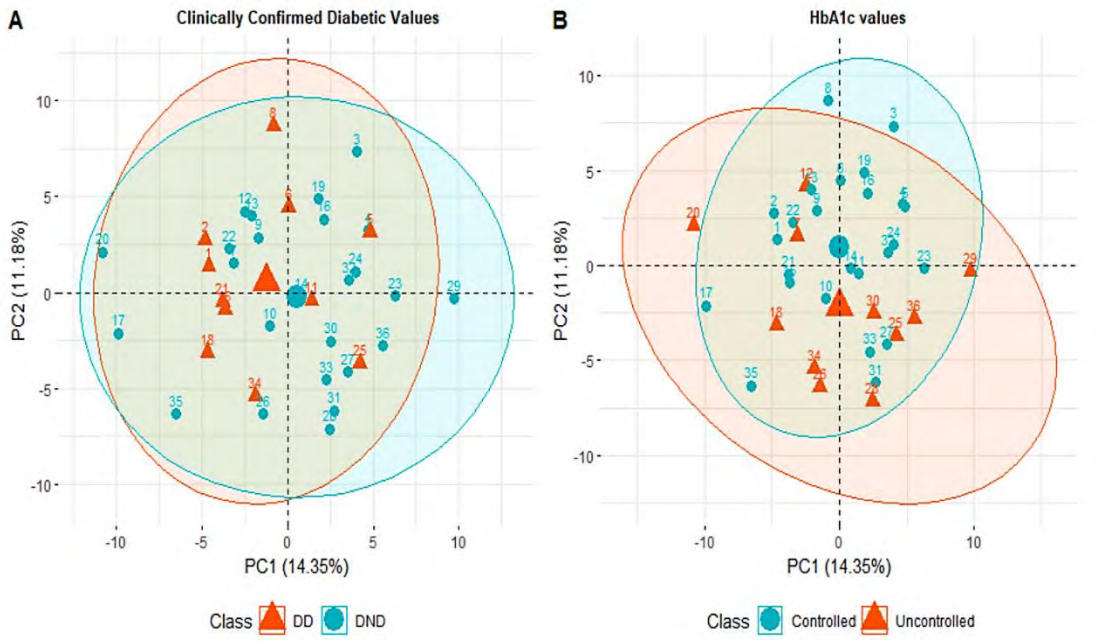


Figure 6-2: Plots of PCA scores. (A) PCA plot based on clinically confirmed diabetic status, (B) PCA plot based on latest HbA1c values.

6.3.4 Discrepancy metabolite analysis

Wilcoxon rank-sum test as a robust non-parametric testing procedure was used to examine the differential metabolites among participants groups under the two analysis scenarios discussed previously. First, we conducted the Wilcoxon rank-sum test for all 142 detected metabolites using clinically confirmed diabetic status. Then, FDR adjusted P-values were obtained. Out of the 142 metabolites, five metabolites significantly had different concentrations among the DD and DND groups. Boxplot of these metabolites intensities are shown in Figure 6-3A with adjusted p-values of Elaidic acid (P = 0.036), Phosphorylcholine (P = 0.036), and Phthalic acid (P = 0.036), the levels of 11a-Hydroxyprogesterone (P = 0.036) and 3,4-Dihydroxymandelic acid (P = 0.036). Analysis was repeated according to the latest HbA1c values as controlled or uncontrolled. Boxplots of the identified significant metabolites according to the HbA1c values are shown in Figure 6-3B. These Boxplots show significant difference in the

levels of Androstenedione (P = 0.042), Delta-hexanolactone (P = 0.042), 2-Furoylglycine (P = 0.042), Maltitol (P = 0.045), Vitamin D3 (P = 0.045), and Indolelactic acid (P = 0.049) and the levels of Isovalerylglycine (P = 0.049).

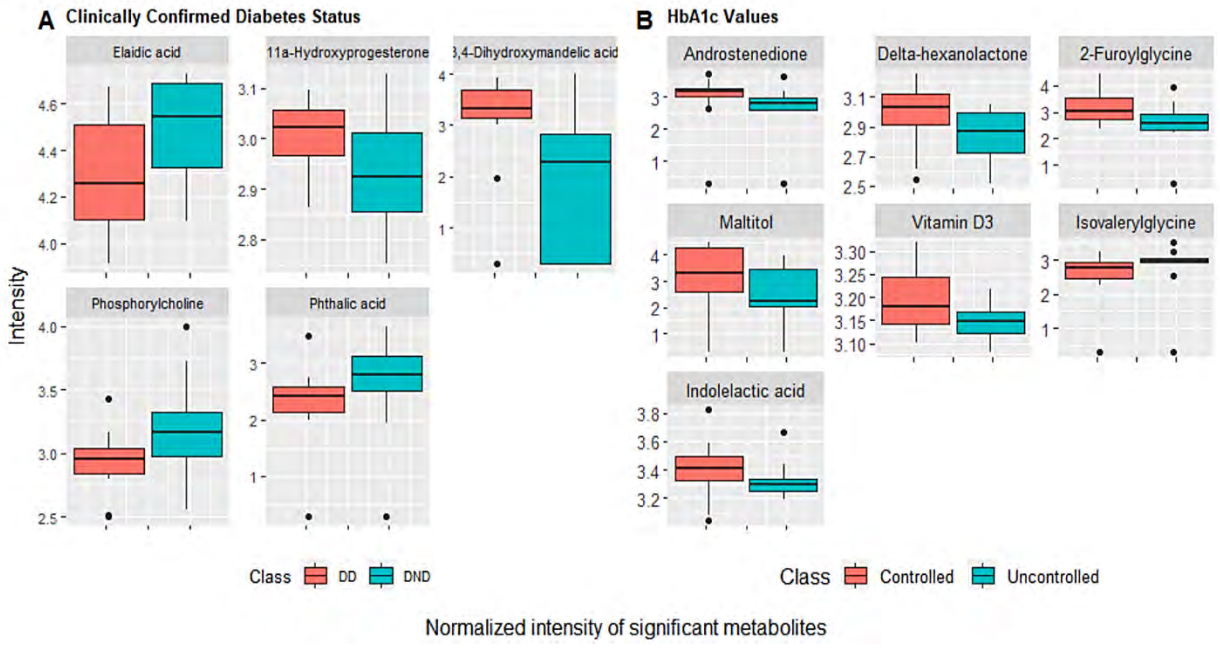


Figure 6-3: (A) Boxplot of normalized intensity metabolites for the clinically confirmed diabetic status. (B) Boxplot of normalized intensity metabolites based on latest HbA1c values.

6.3.5 Analysis of metabolic pathway

First, we examined the PA for the clinically confirmed diabetic status. Based on the identified metabolites in our data, the MetaboAnalyst platform detected 46 metabolic pathways, as exhibited in Figure 6-4A. According to the -log(P) value and pathway impact score, the top three metabolic pathways were selected: Tyrosine metabolism, Linoleic acid metabolism, and Caffeine metabolism. Metabolic PA results show nine different metabolites enriched in these three metabolic pathways: Linoleic acid, Glycerophosphocholine, Paraxanthine, Caffeine, 3,4-Dihydroxymandelic acid, 3,4-Dihydroxyphenylglycol, 3,4-Dihydroxybenzeneacetic acid, DL-Dopa, L-Tyrosine, as shown in Table 6-1. Wilcoxon rank-sum test was used to analyze the differential metabolites enriched in the identified pathways. The levels of Tyrosine metabolism-related metabolite 3,4-Dihydroxymandelic acid (P = 0.028) are noticeably different in both groups. However, there was no significant difference in the levels of other metabolites between the DD group and the DND group. The same approach was applied, considering the grouping of participants based on the latest HbA1c values. In this case, 46 metabolic pathways were screened by the MetaboAnalyst platform. Figure

6-4B shows the top six selected metabolic pathways based on -log(P) value and pathway impact score, which are: Citrate cycle, Glycerolipid metabolism, Vitamin B6 metabolism, Caffeine metabolism, Phenylalanine, tyrosine, tryptophan biosynthesis, and Linoleic acid metabolism. Metabolic PA results indicated 11 different metabolites enriched in these six metabolic pathways: Cis-Aconitic acid, Glycerol, Pyridoxal 5'-phosphate, Pyridoxal, 4-Pyridoxic acid, Caffeine, Paraxanthine, L-Phenylalanine, L-Tyrosine, Linoleic acid, and Glycerophosphocholine, as shown in Table 6-1. Wilcoxon rank-sum test was used to analyze the differential metabolites enriched in the identified metabolism pathways. The levels of glycerolipid metabolism-related metabolite Glycerol (P = 0.050) were significantly different among the controlled and uncontrolled groups. However, there was no significant difference in the levels of other metabolites between the two groups.

Table 6-1: Analysis of the top metabolic pathways based on clinically confirmed diabetic status and latest HbA1c values.

	Name	**-Log(P)**	**Impact**	**Compounds**	**Pathway**
Clinically confirmed diabetic status	Linoleic acid metabolism	0.30064	1.0	Linoleic acid, Glycerophosphocholine	hsa00591
	Caffeine metabolism	1.7512	0.69231	Paraxanthine, Caffeine	map00232
	Tyrosine metabolism	1.7414	0.27636	3,4-Dihydroxymandelic acid, 3,4-Dihydroxyphenylglycol, 3,4-Dihydroxybenzeneacetic acid, DL-Dopa, L-Tyrosine	map00350
Latest HbA1c values	Citrate cycle	1.5898	0.05003	Cis-Aconitic acid	hsa00020
	Glycerolipid metabolism	1.1213	0.23676	Glycerol	hsa00561
	Vitamin B6 metabolism	0.67539	0.68759	Pyridoxal 5'-phosphate, Pyridoxal, 4-Pyridoxic acid	hsa00750
	Linoleic acid metabolism	0.08917	1.0	Linoleic acid, Glycerophosphocholine	hsa00591
	Caffeine metabolism	0.37079	0.69231	Paraxanthine, Caffeine	hsa00232
	Phenylalanine, tyrosine, and tryptophan biosynthesis	0.19682	1	L-Phenylalanine, L-Tyrosine,	hsa00400

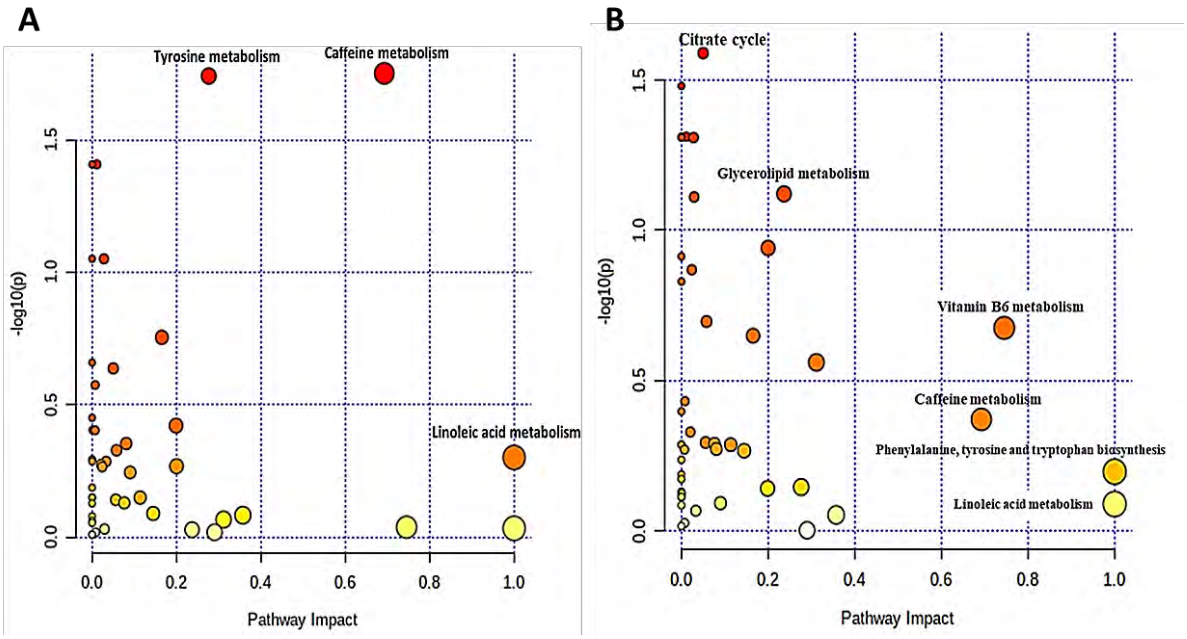


Figure 6-4: (A) Metabolic pathway analysis of Clinically confirmed diabetic status. (B) Metabolic pathway analysis based on latest HbA1c values.

6.4 Discussion and Conclusions

This study used LC-MS/MS to examine blood metabolites of DD and DND Emirati patients. The analysis is two-fold: (1) the analysis is based on clinically confirmed diabetic status, and (2) the analysis is based on the available HbA1c values. We detected and identified 142 metabolites among the DD and DND groups. Initial results using PCA of clinically confirmed diabetic status showed that DD and DND of the plasma components could not have apparent clustering. Therefore, we further performed PCA using HbA1c values. The results showed that the uncontrolled group could be clearly distinguished from the controlled group, indicating that the controlled and uncontrolled groups' plasma metabolites are different.

Subsequently, the Wilcoxon rank-sum test and metabolic PA of 142 metabolites were performed. Differential metabolites analysis based on the clinically confirmed diabetic status between both groups showed enrichment of Hydroxyprogesterone (P = 0.036). It was consistent with previous publications related to androgenic metabolism, oxidative stress, and adipocyte activity accumulation among the DD group [274, 275]. Therefore, the inability to metabolize androstenedione to testosterone and accumulation in blood among the DD group could be correlated with DKD and thereby a useful biomarker. Moreover, we detected an alteration in norepinephrine derivative, 3,4-Dihydroxymandelic acid (P = 0.036) turnover and metabolism among the DD group

and consistent another diabetic sequela such as diabetic cardiomyopathy [276]. Similarly, we identified higher levels of isovalerylglycine (P = 0.049) among the uncontrolled group based on HbA1c values.

Interestingly and consistent with a previous study concluded a higher clearance rate among DKD compared to vascular causes of kidney disease [277]. Vitamin D is an essential regulator of calcium and phosphate homeostasis. Surprisingly, despite the expected decline in kidney function, including 1α-hydroxylation (a necessary step in vitamin D metabolism), we detected an increase in vitamin D3 (P = 0.045) among the controlled group based on HbA1c values [278]. Perhaps due to compensatory mechanisms by other organs such as the gastrointestinal system, for example, a previous study showed a protective role against creatinine degradation among individuals with diabetes with high HbA1c values [279]. Furthermore, plasma metabolites of the Glycerolipid metabolism pathways such as Glycerol (P = 0.05) were increased in the uncontrolled group based on HbA1c values. Interestingly, a previous study concluded that altered tissue lipid metabolism is involved in the pathogenesis of toxin-induced nephropathy and perhaps can be used as an early screening biomarker [280].

Last, mitochondrial dysfunction is one of the mechanisms that contribute to the incidence and development of DKD [281-283]. Mitochondrial dysfunction is associated with kidney disease in non-diabetic contexts, and increasing evidence indicates that dysfunctional renal mitochondria are pathological mediators of DKD [282]. Besides, studies revealed that fatty acid metabolism disorders contribute to the development of DKD in T2DM patients [284]. For example, previous western studies concluded that the lower intake of polyunsaturated fatty acids, primarily linolenic and linoleic acid, is associated with CKD in T2DM patients [285, 286]. Our study found that elaidic acid (P = 0.036) decreased in the DD group. Therefore, targeting key enzymes for such metabolites may be a promising avenue in treating DKD, especially advanced-stage DKD such as ESRD.

We acknowledge the limitation of the small number of patients enrolled in this study. In addition, this one-site study requires a follow-up with a larger cohort to validate our findings further. Furthermore, some of the identified metabolites such as caffeine can

be further attributed to other factors such as diet and medication-the need for further validation.

In conclusion, metabolomics is an emerging technology that plays an essential role in better understanding health and disease conditions as metabolic biomarkers have translational potential to improve disease diagnosis and therapeutic targets. Herein, we identified for the first-time potential biomarkers, such as isovalerylglycine, elaidic acid, hydroxyprogesterone, 3,4-Dihydroxymandelic acid, and glycerolipid metabolites such as Glycerol for early detection of DKD based on robust metabolomics modeling between diabetic hemodialysis and non-diabetic hemodialysis patients in the UAE population.

Chapter 7. A Framework for Optimum Biomarker Discovery

The sophisticated paradigm of metabolomics studies requires researchers to consider each step and employ it thoroughly. However, the abundance of metabolomics studies doesn't necessarily lead to validated outcomes. Moreover, the lack and weaknesses in experiment designs deter medical experts from implementing the results in the clinical setting. However, the ultimate goal for metabolomics studies is to translate the hypothesized knowledge toward better disease therapeutic and management.

Recently, several studies proposed solutions for some of the detected issues in the body of the experiment. Unfortunately, however, the road is longer than expected.

The disintegrated proposed knowledge for reaching optimal outcomes requires an informed and detailed paradigm combining the complete route from biological samples collection to creating a worldwide consensus for biomarkers in each targeted disease. Therefore, we proposed a comprehensive guide for optimal biomarker discovery (Figure 7-1). The recommended framework is motivated by previous studies and based on the challenges we faced in our research. Furthermore, earlier works tackle specific elements of metabolomics studies. Therefore, to the best of our knowledge, our framework is the first to discuss and propose the complete aspects of metabolomics experiments to achieve optimal outcomes.

The very first step in metabolomics studies is biological samples acquisition. This step is a very hectic and challenging part of the whole endeavor. Typically, the quest for an ideal sample is non-realistic. The subject selection is often driven more by specimen availability than a rigorous study protocol. However, sampling selection substantially affects subsequent steps and leads to false discoveries. This chaotic samples collection is subject to massive bias and could limit the complete patient data for support and be incapable of satisfying power calculations based on subject inclusion criteria. This constraint makes the subsequent clinical validity questionable.

As stated before, metabolomics focuses on biomarker discovery to identify metabolites associated with different diseases and environmental exposures. Despite the wealth of studies on metabolomics, the causality assessment is often complex because of confounding, reverse causation, and other uncertainties. Therefore, the caution of inter-individual metabolite variation arising due to differences in genetic factors and

environmental influences should be considered when collecting samples. These stimuli result in several metabolic responses in population studies [287], leading to extreme difficulty locating metabolites related to a specific disorder and, eventually, providing clinical biomarkers [288]. This dilemma is the case, especially when analyzing a multifaceted condition such as diabetes. Several methods can overcome biological variations in human studies. Creating appropriate experimental design and statistical power for the research and using patient questionnaires following population stratification and regression modeling can obtain essential metabolites [288]. These approaches can remove confounding samples from the analysis and help rationalize the data to identify metabolites only correlated with the biological stimulus.

Moreover, metabolite normalization strategies such as evaluating metabolite ratios or normalizing to creatinine in urine experiments could help. Also, a collective effort is required to develop databases collecting data on the normal variations in metabolite concentration ranges in response to influences such as age, gender, diet, pollution, and exercise, which are common reasons for sample-to-sample fluctuation. The last proposal promotes the exposome research [289], measuring all exposures to which an individual is subjected from conception to death and how those exposures relate to health. However, no universal method exists to determine the entirety of the exposome yet [290]. The figure shows that inter-individual genomic, environmental, and gut microflora variation can add to an individual-specific metabotype or metabolomic fingerprint. Metabotype comprises all the genetic, environmental, and gut microflora modifications that are not readily noticeable, and it provides each individual a defining metabolomic fingerprint. Each of these factors can influence the others and determine the outcome of the metabotype. On the other hand, the individual’s metabolome can affect each one of the factors [287].

The most widely employed analytical techniques in metabolomics are MS and NMR. Each analytical approach has its benefits and drawbacks. None of these tools can examine all compounds; combining two or three analytical platforms is required to detect disease pathogenesis and discriminate conditions. The factors that specify the platform selection are the study's focus and samples. More factors limit the platform choice, such as its availability, the cost, and the existing expertise.

These different platforms do not compete, as none of them can conduct a complete detection and quantification of all metabolites set for a targeted biological sample. Accordingly, the optimum metabolomic experiments utilize various technology platforms [37, 235-237] to enhance metabolite annotation and detection. Together NMR and MS are equally crucial for metabolomics experiments. Combining several analytical sources is vital to the future of metabolomics research.

Pathway analysis encompasses different major elements that yield improved outcomes if scientifically employed. However, the metabolomics community lacks a common standard describing its best-practice use. In addition, several parameters in pathway analysis have a pronounced effect on the outcome, yet their impact caught a little systematic attention in the area. These parameters include differential metabolite selection methods, false discovery rate threshold, pathway size, and pathway database choice, generating different results.

Databases are considered the cornerstone in pathway analysis, and recent studies revealed that choosing a pathway database could substantially affect the results [203, 244]. During recent decades, advancements in pathway databases have caused several formalization schemes, impeding the interoperability among these resources and generating data silos. Pathway database selection [291], metabolite misidentification rate [292], and assay chemical bias of several analytical platforms [203] will impact most PA methods. Therefore, the suggestion to overcome the databases pitfalls is to perform organism-specific pathway analysis using multiple pathway databases and form a consensus pathway signature utilizing the outcomes. Databases integration comprising multiple pathway databases, such as the ConsensusPathDB [293] or PathMe [294], might be helpful and consider continuing attempts to standardize pathway resources.

An effort [109] presented a critical overview, for the first time, of the performance of selected bioinformatics tools for omics datasets. These tools include BioCyc/HumanCyc [181], ConsensusPathDB [293], IMPaLA [217], MBRole [209], MetaboAnalyst [211], Metabox [295], MetExplore [207], MPEA [212], PathVisio [296], and Reactome [180] and KEGGREST [228]. Despite the variability of the tools, they generated coherent results independent of their analytical method. Nevertheless,

as indicated before, further effort on the completeness of metabolite and pathway databases is necessary, which dramatically impacts the analysis accuracy.

Integrating biological knowledge and machine learning received significant attention in recent years [297, 298]. With the combination of pathway information and novel statistical methods, machine learning tools represent a promising computational approach in examining pathways and could deliver biological insight into the study of metabolomics data. However, the fundamental challenge confronting the successful application of ML in such data is the essential need for high-quality and large quantity datasets. High repeatability, reproducibility, and minimal uncertainty are crucial aspects of high-quality data. An experiment should generate similar responses using the same inputs; otherwise, there is little promise that an algorithm can be predictive [298]. The above recommendation in the sampling process should enhance ML applications if considered.

The goal of pathway analysis is to attain a list of potential biomarkers for disease diagnosis and prognosis. However, these perturbed biomarkers require necessary validation steps to transfer them to clinical use successfully. The following recommended approach is a targeted metabolomics experiment, which allows validation and absolute quantification of a predefined list of potential metabolites. Targeted metabolomics provides higher sensitivity and selectivity than untargeted metabolomics. Furthermore, targeted metabolomics can optimize sample preparation, decreasing the dominance of high-abundance molecules in the studies. In addition, due to prior knowledge of metabolites of interest, analytical artifacts are not transferred to downstream analysis. Therefore, this approach can partially validate the reproducibility and repetitiveness of the results. The follow–up experiments should be carried out in an additional cohort of biological samples to validate the metabolite variations with specific conditions.

It is reasonable to say that metabolomics studies are concerned primarily with introducing pioneering results within the research community rather than researching the real-time impact of these efforts on health. Therefore, the translations of assays results into practical applications should be considered to accomplish optimum validation. Unfortunately, very few molecular biomarkers are in clinical use [299].

However, many studies acknowledged the existing gap between biomarker discovery and meaningful clinical use [288, 300].

Clinical trials for predefined potential biomarkers are valuable. The parameters that should be tested among these clinical trials [300] include (1) analytical validity to assess reproducibility and accuracy of the test, (2) clinical validity to evaluate a biomarker's ability to distinguish one group from another in a meaningful manner, (3) clinical utility to assess changes of the test results on the outcomes and, (4) evaluating cost-efficient, psychological and ethical implications in case there is value-added or cost saved by knowing the results. Also, is there a possibility of characterizing treatment or risk reduction strategy based on results?

The outcomes of initial clinical trials should be used as feedback to improve the biomarker discovery process, eliminating related challenges, and increasing the benefits.

If each part of the metabolomics experiment is best cured, we can create a trustful global validated biomarkers bank that clinician can quickly embrace.

Multidimensional assessment can provide insight into the systemic biological variations correlated with metabolites and probably direct more mechanistic analyses.

Developing a clinically meaningful biomarker starts with a predefined plan. But, when clinicians and medical professionals are unsure what to do with the abundance of such information, how best can individuals be informed?

Together, scientists have a unique opportunity to abandon malformed research practices and stand together against this field's overwhelming and complex nature in a uniform and scientifically sound approach. Metabolomics can promote and motivate researchers and practitioners to think outside the box and stand as leaders in discovering new avenues to reach better public health, improving clinical care and patient outcomes.

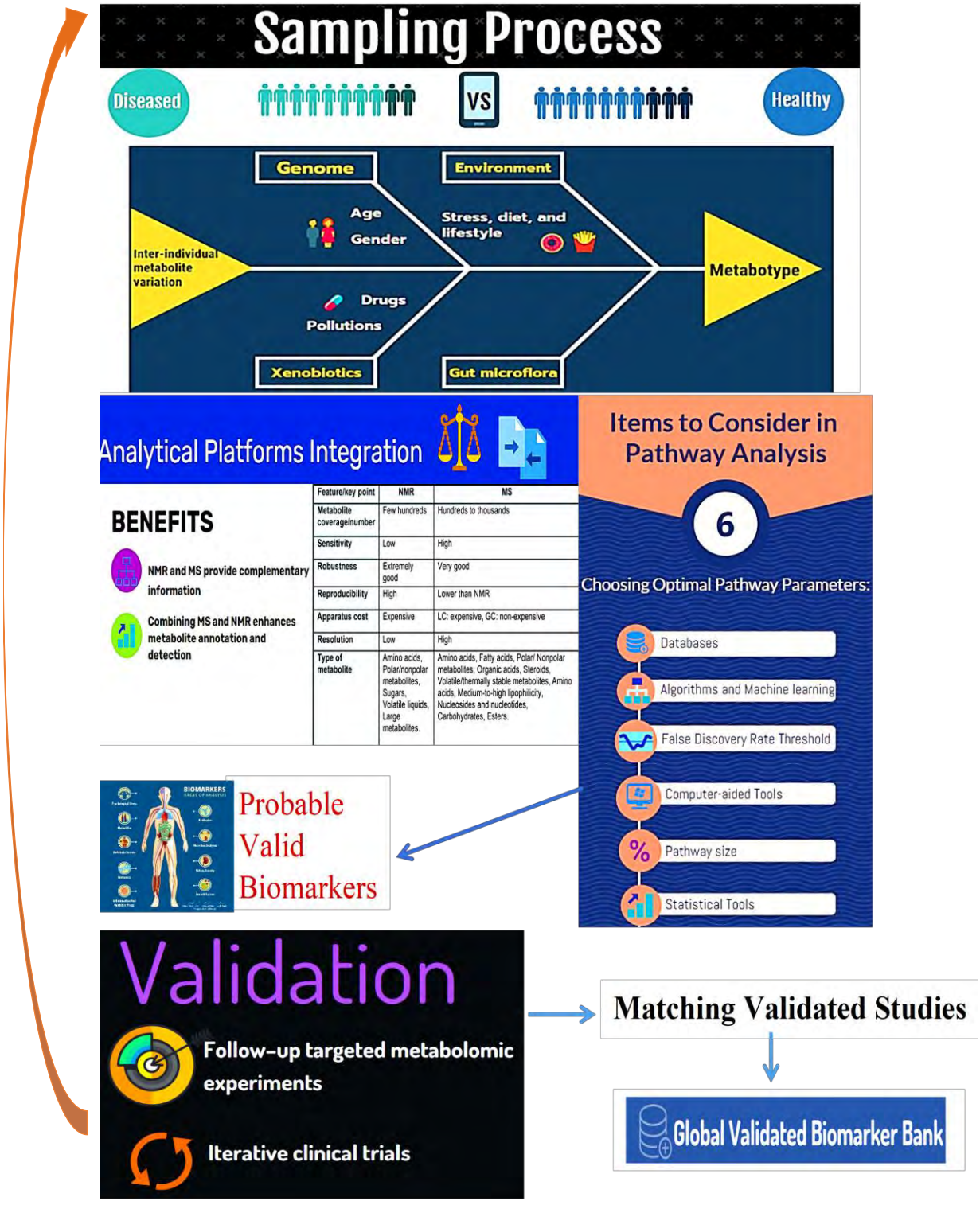


Figure 7-1: Framework for optimal biomarker discovery.

Chapter 8. Concluding Remarks

Good quality health is fundamental to human happiness and well-being, promoting prosperity, wealth, and even economic growth, as healthy populations are more productive and have longevity. Therefore, nations worldwide contribute toward research enriching people's health. As chronic diseases are a significant problem globally, they are continuously and unstoppable to fight them.

The metabolomics experiments stand up as a considerable prospective driver for defining the mechanisms of different diseases. Moreover, the multifaceted nature of such experiments encouraged physicians, chemists, biologists, engineers, and IT professionals to build a robust metabolomics workbench. The intricated journey of metabolomics experiments starts from collecting species and ends with knowledgeable information to boost well-being and resolve related issues.

Diabetes and its related complications are leading causes of morbidity and mortality globally. Therefore, the surge of studies to solve the riddles behind it is remarkable. This study examines metabolomics' role in diabetes globally and utilizes existing tools and knowledge to link triggers of diabetes in UAE citizens to other nations. The importance of the study lies as it is considered the first comprehensive non-targeted metabolomics experiment to study UAE metabolic profile.

This dissertation is twofold to explore diabetics populations. First, we investigated the metabolic profile between diabetics patients against healthy. Second, a study examined the metabolic profile for diabetic versus non-diabetic dialysis patients.

The first study analyzed blood metabolites of T2DM and non-T2DM Emirati patients. Three scenarios were examined to search for differential metabolites: (1) non-diabetics vs. uncontrolled diabetics, (2) non-diabetics vs. prediabetics and controlled diabetics, (3) and lastly, uncontrolled diabetics vs. prediabetics and controlled diabetics. A panel of 41 metabolic signatures was identified for the groupwise comparisons. The identified metabolites are sorted into classes such as Tryptophan and Purines. Adenine levels are higher in prediabetics and managed diabetics than in uncontrolled diabetics and non-diabetics. Furthermore, the study identified fifteen significant differentially abundant metabolites between uncontrolled diabetics against prediabetics and controlled diabetics. Some metabolites are linked to changes in fatty acid, purine, and bile acid

metabolism, as well as Steroids, Glycerophosphocholines, Diarylethers, Phenylacetic acids, and Indoles metabolism. It is worth highlighting that uncontrolled diabetics patients have cortisol excess compared to prediabetics, controlled diabetics, and non-diabetics.

The second study introduced several potential metabolites helping understand dialysis diabetic phenotype. It identified for the first-time potential biomarkers, such as isovalerylglycine, elaidic acid, hydroxyprogesterone, 3,4-Dihydroxymandelic acid, and glycerolipid metabolites such as Glycerol for early detection of DKD based on robust metabolomics modeling between diabetic hemodialysis and non-diabetic hemodialysis patients in the UAE population.

Utilizing big data, computer-aided tools, and established databases and repositories helped create a metabolic starting point for UAE studies in Omics technologies. The study outcomes are enlightening toward UAE goals combating diabetes. The discovered biomarkers would be validated by conducting the following Omics studies in UAE. The clinical translation of novel biomarkers could expedite the treatment process and boost the healthcare system beating increasing numbers of diabetes.

PA is the core of metabolomics experiments providing clues about the mechanism of phenotype. Over the last few years, several PA methods have been suggested. However, despite such analysis's power and substantial potential in diseases definitions, it is still rudimentary and ad hoc. We acknowledge metabolomics assay's limitations and weaknesses, also stated in other studies.

One challenge in PA methods in assessing the correctness of whatever comes out from the PA. Often, articles describe new ways and support them using two to three data sets, followed by results interpretation. However, the approach is subjective and biased. Living organisms are complex systems, and some references will support almost any analysis result. Furthermore, with a lack of knowledge about phenotypes phenomena, it is irrational to conclude accurately whether such associations are meaningful or not. The pathways grow as more knowledge is collected. The knowledge obtained by the pathways is both inadequate and partially inaccurate at any moment in time. Even though PA has pitfalls, researchers need to identify significant pathways in the given phenotype. Thus, extensive benchmarking results will be beneficial despite pathway annotations imperfection at one particular time. On the other hand, the heterogeneity in

the individual experiments the number of genes/metabolites found from different studies under the same condition often vary greatly. This issue creates inconsistency and bias toward certain data sets in the downstream analysis.

This exploratory research is initial research into our hypothetical idea to build a solid foundation in diseases treatments. It lays the groundwork for future studies or determines if a current theory might explain what we observed. Explanatory research is required to demonstrate a cause-and-effect phenomenon, investigating patterns and trends in existing data that haven't been previously considered. In our case, future explanative studies need to explain why specific metabolites differ in particular pathways and how signaling pathways work. Our analysis studied associations amongst groups and respective metabolites, while future explanatory studies should explain causation. It is well-known that correlation does not mean causation.

References

[1] A. Artasensi, A. Pedretti, G. Vistoli, and L. Fumagalli, "Type 2 diabetes mellitus: A review of multi-target drugs," *Molecules,* vol. 25, no. 8, p. 1987, 2020.

[2] H. Sun *et al.*, "IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045," *Diabetes Research and Clinical Practice,* p. 109119, 2021.

[3] C. D. Mathers and D. Loncar, "Projections of global mortality and burden of disease from 2002 to 2030," *PLoS Medicine,* vol. 3, no. 11, p. e442, 2006.

[4] J. H. Yun *et al.*, "Metabolomics profiles associated with HbA1c levels in patients with type 2 diabetes," *Plos One,* vol. 14, no. 11, p. e0224274, 2019.

[5] Y. Sun, H.-Y. Gao, Z.-Y. Fan, Y. He, and Y.-X. Yan, "Metabolomics Signatures in Type 2 Diabetes: A Systematic Review and Integrative Analysis," *The Journal of Clinical Endocrinology & Metabolism,* vol. 105, no. 4, pp. 1000-1008, 2020, doi: 10.1210/clinem/dgz240.

[6] B. Arneth, R. Arneth, and M. Shams, "Metabolomics of type 1 and type 2 diabetes," *International Journal of Molecular Sciences,* vol. 20, no. 10, p. 2467, 2019.

[7] S. R. Khan *et al.*, "The discovery of novel predictive biomarkers and early-stage pathophysiology for the transition from gestational diabetes to type 2 diabetes," *Diabetologia,* vol. 62, no. 4, pp. 687-703, 2019.

[8] P. Romagnani *et al.*, "Chronic kidney disease," *Nature Reviews Disease Primers,* vol. 3, no. 1, p. 17088, 2017/11/23 2017, doi: 10.1038/nrdp.2017.88.

[9] A. M. El Nahas and A. K. Bello, "Chronic kidney disease: the global challenge," *The Lancet,* vol. 365, no. 9456, pp. 331-340, 2005/01/22/ 2005, doi: https://doi.org/10.1016/S0140-6736(05)17789-7.

[10] A. C. Webster, E. V. Nagler, R. L. Morton, and P. Masson, "Chronic kidney disease," *The lancet,* vol. 389, no. 10075, pp. 1238-1252, 2017.

[11] R. Z. Alicic, M. T. Rooney, and K. R. Tuttle, "Diabetic kidney disease: challenges, progress, and possibilities," *Clinical Journal of the American Society of Nephrology,* vol. 12, no. 12, pp. 2032-2045, 2017.

[12] R. Saran, B. Robinson, and K. C. Abbott, "United States Renal Data System: 2016 USRDS Annual Data Report: Epidemiology of Kidney Disease in the United States," *American Journal of Kidney Diseases,* vol. 69, no. 3, p. A4, 2017.

[13] M. A. Niewczas *et al.*, "Circulating modified metabolites and a risk of ESRD in patients with type 1 diabetes and chronic kidney disease," *Diabetes Care,* vol. 40, no. 3, pp. 383-390, 2017.

[14] M. Guthoff *et al.*, "Impact of end-stage renal disease on glucose metabolism—a matched cohort analysis," *Nephrology Dialysis Transplantation,* vol. 32, no. 4, pp. 670-676, 2017.

[15] S. M. Deger, C. D. Ellis, A. Bian, A. Shintani, T. A. Ikizler, and A. M. Hung, "Obesity, diabetes and survival in maintenance hemodialysis patients," *Renal Failure,* vol. 36, no. 4, pp. 546-551, 2014.

[16] M. Vijayan, S. Radhakrishnan, G. Abraham, M. Mathew, K. Sampathkumar, and N. P. Mancha, "Diabetic kidney disease patients on hemodialysis: a retrospective survival analysis across different socioeconomic groups," *NDT Plus,* vol. 9, no. 6, pp. 833-838, 2016.

[17] M. Abe *et al.*, "Is there a “burnt-out diabetes” phenomenon in patients on hemodialysis?," *Diabetes Research and Clinical Practice,* vol. 130, pp. 211-220, 2017.
[18] C. M. Rhee, A. M. Leung, C. P. Kovesdy, K. E. Lynch, G. A. Brent, and K. Kalantar-Zadeh, "Updates on the management of diabetes in dialysis patients," 2014, vol. 27: Wiley Online Library, 2 ed., pp. 135-145.
[19] K. Kalantar-Zadeh, S. F. Derose, S. Nicholas, D. Benner, K. Sharma, and C. P. Kovesdy, "Burnt-out diabetes: impact of chronic kidney disease progression on the natural course of diabetes mellitus," *Journal of Renal Nutrition,* vol. 19, no. 1, pp. 33-37, 2009.
[20] C. P. Kovesdy, J. C. Park, and K. Kalantar-Zadeh, "Glycemic control and burnt-out diabetes in ESRD," 2010, vol. 23: Wiley Online Library, 2 ed., pp. 148-156.
[21] C. P. Kovesdy, K. Sharma, and K. Kalantar-Zadeh, "Glycemic control in diabetic CKD patients: where do we stand?," *American Journal of Kidney Diseases,* vol. 52, no. 4, pp. 766-777, 2008.
[22] J. Park, P. Lertdumrongluk, M. Z. Molnar, C. P. Kovesdy, and K. Kalantar-Zadeh, "Glycemic control in diabetic dialysis patients and the burnt-out diabetes phenomenon," *Current Diabetes Reports,* vol. 12, no. 4, pp. 432-439, 2012.
[23] R. D. Fleischmann *et al.*, "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd," *Science,* vol. 269, no. 5223, pp. 496-512, 1995.
[24] E. S. Lander *et al.*, "Initial sequencing and analysis of the human genome," *Nature*, vol. 412, no. 6846, p.565, 2001.
[25] J. C. Venter *et al.*, "The sequence of the human genome," *Science,* vol. 291, no. 5507, pp. 1304-1351, 2001.
[26] J. N. Weinstein, "‘Omic’and hypothesis-driven research in the molecular pharmacology of cancer," *Current Opinion in Pharmacology,* vol. 2, no. 4, pp. 361-365, 2002.
[27] H. Ge, A. J. M. Walhout, and M. Vidal, "Integrating ‘omic’information: a bridge between genomics and systems biology," *TRENDS in Genetics,* vol. 19, no. 10, pp. 551-560, 2003.
[28] K. Mi *et al.*, "Construction and Analysis of Human Diseases and Metabolites Network," (in eng), *Front Bioeng Biotechnol,* vol. 8, p. 398, 2020, doi: 10.3389/fbioe.2020.00398.
[29] M. V. Schneider and S. Orchard, "Omics technologies, data and bioinformatics principles," in *Bioinformatics for Omics Data*: Springer, 2011, pp. 3-30.
[30] M. Debnath, G. B. K. S. Prasad, and P. S. Bisen, "Omics technology," in *Molecular Diagnostics: Promises and Possibilities*: Springer, 2010, pp. 11-31.
[31] R. R. Egea, N. G. Puchalt, M. M. Escrivá, and A. C. Varghese, "OMICS: current and future perspectives in reproductive medicine and technology," *Journal of Human Reproductive Sciences,* vol. 7, no. 2, p. 73, 2014.
[32] S. C. Gates and C. C. Sweeley, "Quantitative metabolic profiling based on gas chromatography," *Clinical Chemistry,* vol. 24, no. 10, pp. 1663-1673, 1978.
[33] G. G. Harrigan and R. Goodacre, “Metabolic profiling: its role in biomarker discovery and gene function analysis: its role in biomarker discovery and gene function analysis,” *Springer Science & Business Media*, p.335, 2003.
[34] R. Goodacre, S. Vaidyanathan, W. B. Dunn, G. G. Harrigan, and D. B. Kell, "Metabolomics by numbers: acquiring and understanding global metabolite data," *TRENDS in Biotechnology,* vol. 22, no. 5, pp. 245-252, 2004.

[35] M. M. Ulaszewska *et al.*, "Nutrimetabolomics: an integrative action for metabolomic analyses in human nutritional studies," *Molecular Nutrition & Food Research,* vol. 63, no. 1, p. 1800384, 2019.
[36] J. F. Crow, "Unequal by nature: A geneticist's perspective on human differences," *Daedalus,* vol. 131, no. 1, pp. 81-88, 2002.
[37] A. Amberg *et al.*, "NMR and MS methods for metabolomics," in *Drug Safety Evaluation*: Springer, 2017, pp. 229-258.
[38] S. G. Oliver, M. K. Winson, D. B. Kell, and F. Baganz, "Systematic functional analysis of the yeast genome," *Trends in Biotechnology,* vol. 16, no. 9, pp. 373-378, 1998.
[39] B. Daviss, "Growing pains for metabolomics: the newest'omic science is producing results--and more data than researchers know what to do with," *The Scientist,* vol. 19, no. 8, pp. 25-29, 2005.
[40] W. B. Dunn and D. I. Ellis, "Metabolomics: current analytical platforms and methodologies," *TrAC Trends in Analytical Chemistry,* vol. 24, no. 4, pp. 285-294, 2005.
[41] E. C. Horning and M. G. Horning, "Human metabolic profiles obtained by GC and GC/MS," *Journal of Chromatographic Science,* vol. 9, no. 3, pp. 129-140, 1971.
[42] J. van der Greef and A. K. Smilde, "Symbiosis of chemometrics and metabolomics: past, present, and future," *Journal of Chemometrics: A Journal of the Chemometrics Society,* vol. 19, no. 5-7, pp. 376-386, 2005.
[43] L. D. Roberts, A. L. Souza, R. E. Gerszten, and C. B. Clish, "Targeted metabolomics," *Current Protocols in Molecular Biology,* vol. 98, no. 1, pp. 30-2, 2012.
[44] A. Klassen *et al.*, "Metabolomics: Definitions and significance in systems biology," in *Metabolomics: From Fundamentals to Clinical Applications*: Springer, 2017, pp. 3-17.
[45] D. S. Wishart *et al.*, "HMDB 5.0: the Human Metabolome Database for 2022," *Nucleic Acids Research,* vol. 50, no. D1, pp. D622-D631, 2022.
[46] M. Kanehisa, M. Furumichi, Y. Sato, M. Ishiguro-Watanabe, and M. Tanabe, "KEGG: integrating viruses and cellular organisms," *Nucleic Acids Research,* vol. 49, no. D1, pp. D545-D551, 2021, doi: 10.1093/nar/gkaa970.
[47] S. Kim *et al.*, "PubChem in 2021: new data content and improved web interfaces," *Nucleic Acids Research,* vol. 49, no. D1, pp. D1388-D1395, 2021, doi: 10.1093/nar/gkaa971.
[48] C. A. Smith *et al.*, "METLIN: a metabolite mass spectral database," *Therapeutic Drug Monitoring,* vol. 27, no. 6, pp. 747-751, 2005.
[49] H. Horai *et al.*, "MassBank: a public repository for sharing mass spectral data for life sciences," *Journal of Mass Spectrometry,* vol. 45, no. 7, pp. 703-714, 2010.
[50] E. Fahy *et al.*, "Update of the LIPID MAPS comprehensive classification system for lipids," *Journal of Lipid Research,* vol. 50, no. Supplement, pp. S9-S14, 2009.
[51] J. Hastings *et al.*, "ChEBI in 2016: Improved services and an expanding collection of metabolites," *Nucleic Acids Research,* vol. 44, no. D1, pp. D1214-D1219, 2016, doi: 10.1093/nar/gkv1031.
[52] E. L. Ulrich *et al.*, "BioMagResBank." *Nucleic Acids Research*, vol. 36, pp. D402–D408, 2008.

[53] T. Jewison *et al.*, "SMPDB 2.0: big improvements to the Small Molecule Pathway Database," (in eng), *Nucleic Acids Research,* vol. 42, no. Database issue, pp. D478-84, doi: 10.1093/nar/gkt1067, Jan 2014.
[54] A. Alonso, S. Marsal, and A. Julià, "Analytical methods in untargeted metabolomics: state of the art in 2015," *Frontiers in Bioengineering and Biotechnology,* vol. 3, p. 23, 2015.
[55] A. Cambiaghi, M. Ferrario, and M. Masseroli, "Analysis of metabolomic data: tools, current strategies and future challenges for omics data integration," *Briefings in Bioinformatics,* vol. 18, no. 3, pp. 498-510, 2017.
[56] A. Sussulini, "Metabolomics: from fundamentals to clinical applications," *Springer*, vol. 965, 2017.
[57] K. M. Sas, A. Karnovsky, G. Michailidis, and S. Pennathur, "Metabolomics and diabetes: analytical and computational approaches," *Diabetes,* vol. 64, no. 3, pp. 718-732, 2015.
[58] A. Hubel, R. Spindler, and A. P. N. Skubitz, "Storage of human biospecimens: selection of the optimal storage temperature," *Biopreservation and Biobanking,* vol. 12, no. 3, pp. 165-175, 2014.
[59] X. Liu and J. W. Locasale, "Metabolomics: a primer," *Trends in Biochemical Sciences,* vol. 42, no. 4, pp. 274-284, 2017.
[60] Z. Pan and D. Raftery, "Comparing and combining NMR spectroscopy and mass spectrometry in metabolomics," (in eng), *Analytical and Bioanalytical Chemistry,* vol. 387, no. 2, pp. 525-7, Jan 2007, doi: 10.1007/s00216-006-0687-8.
[61] E. C. Considine, G. Thomas, A. L. Boulesteix, A. S. Khashan, and L. C. Kenny, "Critical review of reporting of the data analysis step in metabolomics," (in eng), *Metabolomics,* vol. 14, no. 1, p. 7, 12 2017, doi: 10.1007/s11306-017-1299-3.
[62] A. G. Lazar, F. Romanciuc, M. A. Socaciu, and C. Socaciu, "Bioinformatics tools for metabolomic data processing and analysis using untargeted liquid chromatography coupled with mass spectrometry," *Bulletin of University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca. Animal Science and Biotechnologies,* vol. 72, no. 2, pp. 103-115, 2015.
[63] K. H. Liland, "Multivariate methods in metabolomics–from pre-processing to dimension reduction and statistical analysis," *TrAC Trends in Analytical Chemistry,* vol. 30, no. 6, pp. 827-841, 2011.
[64] Z. Amin Mohsin, A. Paul, and S. Devendra, "Pitfalls of using HbA1c in the diagnosis and monitoring of diabetes," *London Journal of Primary Care,* vol. 7, no. 4, pp. 66-69, 2015.
[65] B. Dorcely *et al.*, "Novel biomarkers for prediabetes, diabetes, and associated complications," *Diabetes, Metabolic Syndrome and Obesity:Targets and Therapy,* vol. 10, p. 345, 2017.
[66] M. S. Klein and J. Shearer, "Metabolomics and type 2 diabetes: translating basic research into clinical application," *Journal of Diabetes Research,* vol. 2016, 2016.
[67] S. A. Varvel *et al.*, "Comprehensive biomarker testing of glycemia, insulin resistance, and beta cell function has greater sensitivity to detect diabetes risk than fasting glucose and HbA1c and is associated with improved glycemic control in clinical practice," *Journal of Cardiovascular Translational Research,* vol. 7, no. 6, pp. 597-606, 2014.

[68] Á. López-López, Á. López-Gonzálvez, T. C. Barker-Tejeda, and C. Barbas, "A review of validated biomarkers obtained through metabolomics," *Expert Review of Molecular Diagnostics,* vol. 18, no. 6, pp. 557-575, 2018.
[69] E. Nalejska, E. Mączyńska, and M. A. Lewandowska, "Prognostic and predictive biomarkers: tools in personalized oncology," *Molecular Diagnosis & Therapy,* vol. 18, no. 3, pp. 273-284, 2014.
[70] A. Italiano, "Prognostic or predictive? It's time to get back to definitions," *Journal Clinical Oncology,* vol. 29, no. 35, p. 4718, 2011.
[71] T. J. Lyons and A. Basu, "Biomarkers in diabetes: hemoglobin A1c, vascular and tissue markers," (in eng), *Translational Research,* vol. 159, no. 4, pp. 303-12, Apr 2012, doi: 10.1016/j.trsl.2012.01.009.
[72] D. Ford, D. F. Easton, D. T. Bishop, S. A. Narod, and D. E. Goldgar, "Risks of cancer in BRCA1-mutation carriers. Breast Cancer Linkage Consortium," (in eng), *Lancet,* vol. 343, no. 8899, pp. 692-5, Mar 1994, doi: 10.1016/s0140-6736(94)91578-4.
[73] E. J. Gallagher, D. Le Roith, and Z. Bloomgarden, "Review of hemoglobin A1c in the management of diabetes," *Journal of Diabetes,* vol. 1, no. 1, pp. 9-17, 2009.
[74] Y. Homma, "Predictors of atherosclerosis," *Journal of Atherosclerosis and Thrombosis,* vol. 11, no. 5, pp. 265-270, 2004.
[75] C. S. D. Cruz, L. T. Tanoue, and R. A. Matthay, "Lung cancer: epidemiology, etiology, and prevention," *Clinics in Chest Medicine,* vol. 32, no. 4, pp. 605-644, 2011.
[76] S. M. Aghaei Zarch, M. Dehghan Tezerjani, M. Talebi, and M. Y. Vahidi Mehrjardi, "Molecular biomarkers in diabetes mellitus (DM)," (in eng), *Medical Journal of the Islamic Republic of Iran,* vol. 34, pp. 28-28, 2020, doi: 10.34171/mjiri.34.28.
[77] B. Dorcely *et al.*, "Novel biomarkers for prediabetes, diabetes, and associated complications," (in eng), *Diabetes, Metabolic Syndrome and Obesity : Targets and Therapy,* vol. 10, pp. 345-361, 2017, doi: 10.2147/DMSO.S100074.
[78] J. Xia, D. I. Broadhurst, M. Wilson, and D. S. Wishart, "Translational biomarker discovery in clinical metabolomics: an introductory tutorial," *Metabolomics,* vol. 9, no. 2, pp. 280-299, 2013.
[79] K. Jee and G. H. Kim, "Potentiality of big data in the medical sector: focus on how to reshape the healthcare system," (in eng), *Healthcare Informatics Research,* vol. 19, no. 2, pp. 79-85, Jun 2013, doi: 10.4258/hir.2013.19.2.79.
[80] I. Izonin and N. Shakhovska, "Informatics & data-driven medicine," *Mathematical Biosciences and Engineering,* vol. 18, no. 5, pp. 6430-6433, 2021.
[81] R. M. J. J. van der Kleij *et al.*, "SERIES: eHealth in primary care. Part 1: Concepts, conditions and challenges," *European Journal of General Practice,* vol. 25, no. 4, pp. 179-189, 2019.
[82] M. S. Marcolino, J. A. Q. Oliveira, M. D'Agostino, A. L. Ribeiro, M. B. M. Alkmim, and D. Novillo-Ortiz, "The impact of mHealth interventions: systematic review of systematic reviews," *JMIR mHealth and uHealth,* vol. 6, no. 1, p. e8873, 2018.
[83] S. C. Mathews, M. J. McShea, C. L. Hanley, A. Ravitz, A. B. Labrique, and A. B. Cohen, "Digital health: a path to validation," *NPJ Digital Medicine,* vol. 2, no. 1, pp. 1-9, 2019.

[84] D. T. Patel, "Big Data Analytics in Bioinformatics," in *Biotechnology: Concepts, Methodologies, Tools, and Applications*: IGI Global, 2019, pp. 1967-1984.
[85] A. D. Baxevanis, G. D. Bader, and D. S. Wishart, "Bioinformatics*," John Wiley & Sons*, 2020.
[86] O. World Health, "Genomics and world health: Report of the Advisory Committee on Health Research," *World Health Organization*, 2002.
[87] J. N. Weinstein *et al.*, "The cancer genome atlas pan-cancer analysis project," *Nature Genetics,* vol. 45, no. 10, p. 1113, 2013.
[88] I. Dunham *et al.*, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 487, p. 57, 2012.
[89] F. S. Collins and H. Varmus, "A new initiative on precision medicine," *New England Journal of Medicine,* vol. 372, no. 9, pp. 793-795, 2015.
[90] N. V. Chawla and D. A. Davis, "Bringing big data to personalized healthcare: a patient-centered framework," *Journal of General Internal Medicine,* vol. 28, no. 3, pp. 660-665, 2013.
[91] E. M. Antman *et al.*, "Acquisition, analysis, and sharing of data in 2015 and beyond: a survey of the landscape: a conference report from the American Heart Association Data Summit 2015," *Journal of the American Heart Association,* vol. 4, no. 11, p. e002810, 2015.
[92] K. Donsa, S. Spat, P. Beck, T. R. Pieber, and A. Holzinger, "Towards personalization of diabetes therapy using computerized decision support and machine learning: some open problems and challenges," in *Smart Health*: Springer, 2015, pp. 237-260.
[93] P. Zimmet, K. G. Alberti, D. J. Magliano, and P. H. Bennett, "Diabetes mellitus statistics on prevalence and mortality: facts and fallacies," *Nature Reviews Endocrinology,* vol. 12, no. 10, p. 616, 2016.
[94] P. Saeedi *et al.*, "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas," *Diabetes Research and Clinical Practice,* vol. 157, p. 107843, 2019.
[95] N. R. Hill *et al.*, "Global prevalence of chronic kidney disease–a systematic review and meta-analysis," *PloS One,* vol. 11, no. 7, p. e0158765, 2016.
[96] C.-W. Yang *et al.*, "Global case studies for chronic kidney disease/end-stage kidney disease care," (in eng), *Kidney International Supplements,* vol. 10, no. 1, pp. e24-e48, 2020, doi: 10.1016/j.kisu.2019.11.010.
[97] S. Al-Shamsi, D. Regmi, and R. D. Govender, "Chronic kidney disease in patients at high risk of cardiovascular disease in the United Arab Emirates: A population-based study," (in eng), *PloS One,* vol. 13, no. 6, pp. e0199920-e0199920, 2018, doi: 10.1371/journal.pone.0199920.
[98] S. Barnes *et al.*, "Training in metabolomics research. II. Processing and statistical analysis of metabolomics data, metabolite identification, pathway analysis, applications of metabolomics and its future," (in eng), *J Mass Spectrom,* vol. 51, no. 8, pp. 535-548, Aug 2016, doi: 10.1002/jms.3780.
[99] R. Spicer, R. M. Salek, P. Moreno, D. Cañueto, and C. Steinbeck, "Navigating freely-available software tools for metabolomics analysis," (in eng), *Metabolomics,* vol. 13, no. 9, p. 106, 2017, doi: 10.1007/s11306-017-1242-7.
[100] R. A. van den Berg, H. C. Hoefsloot, J. A. Westerhuis, A. K. Smilde, and M. J. van der Werf, "Centering, scaling, and transformations: improving the

biological information content of metabolomics data," (in eng), *BMC Genomics,* vol. 7, p. 142, Jun 8 2006, doi: 10.1186/1471-2164-7-142.
[101] Å. Rinnan, F. Berg, and S. Engelsen, "Review of the Most Common pre-Processing Techniques for Near-Infrared Spectra," *TrAC Trends in Analytical Chemistry,* vol. 28, pp. 1201-1222, 11/01 2009.
[102] N.-P. V. Nielsen, J. M. Carstensen, and J. Smedsgaard, "Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping," *Journal of Chromatography A,* vol. 805, no. 1, pp. 17-35, 1998/05/01/ 1998.
[103] F. Savorani, G. Tomasi, and S. B. Engelsen, "icoshift: A versatile tool for the rapid alignment of 1D NMR spectra," *Journal of Magnetic Resonance,* vol. 202, no. 2, pp. 190-202, 2010/02/01/ 2010.
[104] Y. Xi and D. Rocke, "Baseline Correction for NMR Spectroscopic Metabolomics Data Analysis," *BMC Bioinformatics,* vol. 9, p. 324, 02/01 2008, doi: 10.1186/1471-2105-9-324.
[105] O. Cloarec *et al.*, "Evaluation of the orthogonal projection on latent structure model limitations caused by chemical shift variability and improved visualization of biomarker changes in 1H NMR spectroscopic metabonomic studies," *Analytical Chemistry,* vol. 77, no. 2, pp. 517-526, 2005.
[106] O. Cloarec *et al.*, "Statistical total correlation spectroscopy: an exploratory approach for latent biomarker identification from metabolic 1H NMR data sets," *Analytical Chemistry,* vol. 77, no. 5, pp. 1282-1289, 2005.
[107] M. A. García-Campos, J. Espinal-Enríquez, and E. Hernández-Lemus, "Pathway analysis: state of the art," *Frontiers in Physiology,* vol. 6, p. 383, 2015.
[108] O. Folger, L. Jerby, C. Frezza, E. Gottlieb, E. Ruppin, and T. Shlomi, "Predicting selective drug targets in cancer through metabolic networks," *Molecular Systems Biology,* vol. 7, no. 1, p. 501, 2011.
[109] A. Marco-Ramell *et al.*, "Evaluation and comparison of bioinformatic tools for the enrichment analysis of metabolomics data," *BMC Bioinformatics,* vol. 19, no. 1, pp. 1-11, 2018.
[110] P. Khatri, M. Sirota, and A. J. Butte, "Ten years of pathway analysis: current approaches and outstanding challenges," *PLoS Computational Biology,* vol. 8, no. 2, p. e1002375, 2012.
[111] A.-L. Barabasi and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nature Reviews Genetics,* vol. 5, no. 2, pp. 101-113, 2004.
[112] Y. Zhu, C. Song, and H. Huo, "Research of traditional Chinese medicine in treating type 2 diabetes and insulin resistance," *World Chinese Medicine,* pp. 135-137, 2015.
[113] P. P. Devarshi, S. M. McNabney, and T. M. Henagan, "Skeletal muscle nucleo-mitochondrial crosstalk in obesity and type 2 diabetes," *International Journal of Molecular Sciences,* vol. 18, no. 4, p. 831, 2017.
[114] J. Gao, C. Duan, and L. J. Li, "The pathogenesis mechanisms of type 2 diabetes mellitus," *Medical Recapitulate,* vol. 21, pp. 3935-3938, 2015.
[115] R. H. Unger and L. Orci, "Glucagon and the A cell: physiology and pathophysiology," *New England Journal of Medicine,* vol. 304, no. 26, pp. 1575-1580, 1981.
[116] M. T. Goodarzi, A. A. Navidi, M. Rezaei, and H. Babahmadi-Rezaei, "Oxidative damage to DNA and lipids: correlation with protein glycation in

patients with type 1 diabetes," *Journal of Clinical Laboratory Analysis,* vol. 24, no. 2, pp. 72-76, 2010.
[117] P. L. Hooper, G. Balogh, E. Rivas, K. Kavanagh, and L. Vigh, "The importance of the cellular stress response in the pathogenesis and treatment of type 2 diabetes," *Cell Stress and Chaperones,* vol. 19, no. 4, pp. 447-464, 2014.
[118] M. Urbanová and M. Haluzik, "The role of adipose tissue in pathogenesis of type 2 diabetes mellitus," *Ceskoslovenska Fysiologie,* vol. 64, no. 2, pp. 73-78, 2015.
[119] D. M. Erion, H.-J. Park, and H.-Y. Lee, "The role of lipids in the pathogenesis and treatment of type 2 diabetes and associated co-morbidities," *BMB Reports,* vol. 49, no. 3, p. 139, 2016.
[120] Q. Ma *et al.*, "Progress in metabonomics of type 2 diabetes mellitus," *Molecules,* vol. 23, no. 7, p. 1834, 2018.
[121] P. Felig, E. Marliss, and G. F. Cahill Jr, "Plasma amino acid levels and insulin secretion in obesity," *New England Journal of Medicine,* vol. 281, no. 15, pp. 811-816, 1969.
[122] J. A. Luetscher, "The metabolism of amino acids in diabetes mellitus," *The Journal of Clinical Investigation,* vol. 21, no. 3, pp. 275-279, 1942.
[123] W. B. Kannel, "Lipids, diabetes, and coronary heart disease: insights from the Framingham Study," *American Heart Journal,* vol. 110, no. 5, pp. 1100-1107, 1985.
[124] A. D. Mooradian, "Dyslipidemia in type 2 diabetes mellitus," *Nature Reviews Endocrinology,* vol. 5, no. 3, pp. 150-159, 2009.
[125] M. R. Taskinen, "Diabetic dyslipidaemia: from basic research to clinical practice," *Diabetologia,* vol. 46, no. 6, pp. 733-749, 2003.
[126] A. Hameed, P. Mojsak, A. Buczynska, H. A. R. Suleria, A. Kretowski, and M. Ciborowski, "Altered metabolome of lipids and amino acids species: a source of early signature biomarkers of T2DM," *Journal of Clinical Medicine,* vol. 9, no. 7, p. 2257, 2020.
[127] T. Wu, S. Qiao, C. Shi, S. Wang, and G. Ji, "Metabolomics window into diabetic complications," *Journal of Diabetes Investigation,* vol. 9, no. 2, pp. 244-255, 2018.
[128] M. Guasch-Ferré *et al.*, "Metabolomics in prediabetes and diabetes: a systematic review and meta-analysis," *Diabetes Care,* vol. 39, no. 5, pp. 833-846, 2016.
[129] Y. Lu *et al.*, "Metabolic signatures and risk of type 2 diabetes in a Chinese population: an untargeted metabolomics study using both LC-MS and GC-MS," *Diabetologia,* vol. 59, no. 11, pp. 2349-2359, 2016.
[130] K. Suhre *et al.*, "Metabolic footprint of diabetes: a multiplatform metabolomics study in an epidemiological setting," *PloS one,* vol. 5, no. 11, p. e13953, 2010.
[131] R. Wang-Sattler *et al.*, "Novel biomarkers for pre-diabetes identified by metabolomics," *Molecular Systems Biology,* vol. 8, no. 1, p. 615, 2012.
[132] A. Floegel *et al.*, "Identification of serum metabolites associated with risk of type 2 diabetes using a targeted metabolomic approach," *Diabetes,* vol. 62, no. 2, pp. 639-648, 2013.
[133] C. Menni *et al.*, "Biomarkers for type 2 diabetes and impaired fasting glucose using a nontargeted metabolomics approach," *Diabetes,* vol. 62, no. 12, pp. 4270-4276, 2013.
[134] F. Xu, S. Tavintharan, C. F. Sum, K. Woon, S. C. Lim, and C. N. Ong, "Metabolic signature shift in type 2 diabetes mellitus revealed by mass

spectrometry-based metabolomics," *The Journal of Clinical Endocrinology & Metabolism,* vol. 98, no. 6, pp. E1060-E1065, 2013.
[135] A. Mastrangelo *et al.*, "Insulin resistance in prepubertal obese children correlates with sex-dependent early onset metabolomic alterations," *International Journal of Obesity,* vol. 40, no. 10, pp. 1494-1502, 2016.
[136] X. Liu *et al.*, "Identification of metabolic biomarkers in patients with type 2 diabetic coronary heart diseases based on metabolomic approach," *Scientific Reports,* vol. 6, no. 1, pp. 1-13, 2016.
[137] R. Pallares-Méndez, C. A. Aguilar-Salinas, I. Cruz-Bautista, and L. del Bosque-Plata, "Metabolomics in diabetes, a review," *Annals of Medicine,* vol. 48, no. 1-2, pp. 89-102, 2016.
[138] Z. Y. Tam *et al.*, "Metabolite profiling in identifying metabolic biomarkers in older people with late-onset type 2 diabetes mellitus," *Scientific Reports,* vol. 7, no. 1, pp. 1-12, 2017.
[139] J. Liu *et al.*, "Metabolomics based markers predict type 2 diabetes in a 14-year follow-up study," *Metabolomics,* vol. 13, no. 9, pp. 1-11, 2017.
[140] E. Shokry, A. E. de Oliveira, M. A. G. Avelino, M. M. de Deus, and N. R. Antoniosi Filho, "Earwax: A neglected body secretion or a step ahead in clinical diagnosis? A pilot study," *Journal of Proteomics,* vol. 159, pp. 92-101, 2017.
[141] L. Shi *et al.*, "Plasma metabolites associated with type 2 diabetes in a Swedish population: a case–control study nested in a prospective cohort," *Diabetologia,* vol. 61, no. 4, pp. 849-861, 2018.
[142] Y. Lu *et al.*, "Serum lipids in association with type 2 diabetes risk and prevalence in a Chinese population," *The Journal of Clinical Endocrinology & Metabolism,* vol. 103, no. 2, pp. 671-680, 2018.
[143] S. J. Yang, S.-Y. Kwak, G. Jo, T.-J. Song, and M.-J. Shin, "Serum metabolite profile associated with incident type 2 diabetes in Koreans: findings from the Korean Genome and Epidemiology Study," *Scientific Reports,* vol. 8, no. 1, pp. 1-10, 2018.
[144] C. M. Rebholz *et al.*, "Serum metabolomic profile of incident diabetes," *Diabetologia,* vol. 61, no. 5, pp. 1046-1054, 2018.
[145] A. V. Ahola-Olli *et al.*, "Circulating metabolites and the risk of type 2 diabetes: a prospective study of 11,896 young adults from four Finnish cohorts," *Diabetologia,* vol. 62, no. 12, pp. 2298-2309, 2019.
[146] X. Gu *et al.*, "Distinctive metabolomics patterns associated with insulin resistance and type 2 diabetes mellitus," *Frontiers in Molecular Biosciences,* vol. 7, p. 411, 2020.
[147] S. Salihovic *et al.*, "Non-targeted urine metabolomics and associations with prevalent and incident type 2 diabetes," *Scientific Reports,* vol. 10, no. 1, pp. 1-9, 2020.
[148] G. Satheesh, S. Ramachandran, and A. Jaleel, "Metabolomics-based prospective studies and prediction of Type 2 Diabetes Mellitus risks," *Metabolic Syndrome and Related Disorders,* vol. 18, no. 1, pp. 1-9, 2020.
[149] A. M. Fikri, R. Smyth, V. Kumar, Z. Al-Abadla, S. Abusnana, and M. R. Munday, "Pre-diagnostic biomarkers of type 2 diabetes identified in the UAE's obese national population using targeted metabolomics," *Scientific Reports,* vol. 10, no. 1, pp. 1-10, 2020.
[150] M. Urpi-Sarda, E. Almanza-Aguilera, S. Tulipani, F. J. Tinahones, J. Salas-Salvadó, and C. Andres-Lacueva, "Metabolomics for biomarkers of type 2

diabetes mellitus: advances and nutritional intervention trends," *Current Cardiovascular Risk Reports,* vol. 9, no. 3, p. 12, 2015.
[151] A. Gonzalez-Franquesa, A. M. Burkart, E. Isganaitis, and M.-E. Patti, "What have metabolomics approaches taught us about type 2 diabetes?," *Current Diabetes Reports,* vol. 16, no. 8, pp. 1-10, 2016.
[152] L. W. Sumner *et al.*, "Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI)," (in eng), *Metabolomics : Official Journal of the Metabolomic Society,* vol. 3, no. 3, pp. 211-221, 2007, doi: 10.1007/s11306-007-0082-2.
[153] I. F. Duarte, S. O. Diaz, and A. M. Gil, "NMR metabolomics of human blood and urine in disease research," (in eng), *J Pharm Biomed Anal,* vol. 93, pp. 17-26, May 2014, doi: 10.1016/j.jpba.2013.09.025.
[154] P. Rocca-Serra *et al.*, "Data standards can boost metabolomics research, and if there is a will, there is a way,"*Metabolomics,* vol. 12, pp. 14-14, 2016.
[155] N. Hoffmann *et al.*, "mzTab-M: A Data Standard for Sharing Quantitative Results in Mass Spectrometry Metabolomics," *Analytical Chemistry,* vol. 91, no. 5, pp. 3302-3310, 2019/03/05 2019, doi: 10.1021/acs.analchem.8b04310.
[156] D. K. Trivedi, K. A. Hollywood, and R. Goodacre, "Metabolomics for the masses: The future of metabolomics in a personalized world," (in eng), *New Horiz Transl Med,* vol. 3, no. 6, pp. 294-305, Mar 2017, doi: 10.1016/j.nhtm.2017.06.001.
[157] V. Tolstikov, A. J. Moser, R. Sarangarajan, N. R. Narain, and M. A. Kiebish, "Current Status of Metabolomic Biomarker Discovery: Impact of Study Design and Demographic Characteristics," (in eng), *Metabolites,* vol. 10, no. 6, p. 224, 2020, doi: 10.3390/metabo10060224.
[158] Z. Yu *et al.*, "Human serum metabolic profiles are age dependent," (in eng), *Aging Cell,* vol. 11, no. 6, pp. 960-7, Dec 2012, doi: 10.1111/j.1474-9726.2012.00865.x.
[159] C. Menni *et al.*, "Metabolomic markers reveal novel pathways of ageing and early development in human populations," (in eng), *Int J Epidemiol,* vol. 42, no. 4, pp. 1111-9, Aug 2013, doi: 10.1093/ije/dyt094.
[160] K. Mittelstrass *et al.*, "Discovery of sexual dimorphisms in metabolic and genetic biomarkers," (in eng), *PLoS Genet,* vol. 7, no. 8, p. e1002215, Aug 2011, doi: 10.1371/journal.pgen.1002215.
[161] J. Krumsiek *et al.*, "Gender-specific pathway differences in the human serum metabolome," (in eng), *Metabolomics,* vol. 11, no. 6, pp. 1815-1833, 2015, doi: 10.1007/s11306-015-0829-0.
[162] M. J. Rist *et al.*, "Metabolite patterns predicting sex and age in participants of the Karlsruhe Metabolomics and Nutrition (KarMeN) study," (in eng), *PLoS One,* vol. 12, no. 8, p. e0183228, 2017, doi: 10.1371/journal.pone.0183228.
[163] W. B. Dunn *et al.*, "Molecular phenotyping of a UK population: defining the human serum metabolome," (in eng), *Metabolomics,* vol. 11, pp. 9-26, 2015, doi: 10.1007/s11306-014-0707-1.
[164] R. Chaleckis, I. Murakami, J. Takada, H. Kondoh, and M. Yanagida, "Individual variability in human blood metabolites identifies age-related differences," (in eng), *Proc Natl Acad Sci U S A,* vol. 113, no. 16, pp. 4252-9, Apr 19 2016, doi: 10.1073/pnas.1603023113.
[165] M. W. K. Wong *et al.*, "Plasma lipidome variation during the second half of the human lifespan is associated with age and sex but minimally with BMI," (in eng), *PLoS One,* vol. 14, no. 3, p. e0214141, 2019.

[166] E. Pujos-Guillot *et al.*, "Identification of Pre-frailty Sub-Phenotypes in Elderly Using Metabolomics," (in eng), *Frontiers in Physiology,* vol. 9, pp. 1903-1903, 2019, doi: 10.3389/fphys.2018.01903.
[167] V.-P. Mäkinen and M. Ala-Korpela, "Metabolomics of aging requires large-scale longitudinal studies with replication," (in eng), *Proceedings of the National Academy of Sciences of the United States of America,* vol. 113, no. 25, pp. E3470-E3470, 2016, doi: 10.1073/pnas.1607062113.
[168] T. Niccoli and L. Partridge, "Ageing as a risk factor for disease," (in eng), *Curr Biol,* vol. 22, no. 17, pp. R741-52, Sep 11 2012, doi: 10.1016/j.cub.2012.07.024.
[169] B. F. Darst, R. L. Koscik, K. J. Hogan, S. C. Johnson, and C. D. Engelman, "Longitudinal plasma metabolomics of aging and sex," (in eng), *Aging (Albany NY),* vol. 11, no. 4, pp. 1262-1282, Feb 24 2019, doi: 10.18632/aging.101837.
[170] M. Gonzalez-Freire *et al.*, "Targeted Metabolomics Shows Low Plasma Lysophosphatidylcholine 18:2 Predicts Greater Decline of Gait Speed in Older Adults: The Baltimore Longitudinal Study of Aging," (in eng), *J Gerontol A Biol Sci Med Sci,* vol. 74, no. 1, pp. 62-67, Jan 1 2019, doi: 10.1093/gerona/gly100.
[171] E. P. Rhee, "Metabolomics and renal disease," *Current Opinion in Nephrology and Hypertension,* vol. 24, no. 4, p. 371, 2015.
[172] M. Darshi, B. Van Espen, and K. Sharma, "Metabolomics in Diabetic Kidney Disease: Unraveling the Biochemistry of a Silent Killer," (in eng), *Am J Nephrol,* vol. 44, no. 2, pp. 92-103, 2016, doi: 10.1159/000447954.
[173] C.-J. Chen, W.-L. Liao, C.-T. Chang, Y.-N. Lin, and F.-J. Tsai, "Identification of urinary metabolite biomarkers of type 2 diabetes nephropathy using an untargeted metabolomic approach," *Journal of Proteome Research,* vol. 17, no. 11, pp. 3997-4007, 2018.
[174] X. Tang, J. You, D. Liu, M. Xia, L. He, and H. Liu, "5-Hydroxyhexanoic acid predicts early renal functional decline in type 2 diabetes patients with microalbuminuria," *Kidney and Blood Pressure Research,* vol. 44, no. 2, pp. 245-263, 2019.
[175] H. M. Colhoun and M. L. Marcovecchio, "Biomarkers of diabetic kidney disease," *Diabetologia,* vol. 61, no. 5, pp. 996-1011, 2018.
[176] H. Zhang *et al.*, "Identification of potential serum metabolic biomarkers of diabetic kidney disease: a widely targeted metabolomics study," *Journal of Diabetes Research,* vol. 2020, 2020.
[177] N. Dincer, T. Dagel, B. Afsar, A. Covic, A. Ortiz, and M. Kanbay, "The effect of chronic kidney disease on lipid metabolism," (in eng), *Int Urol Nephrol,* vol. 51, no. 2, pp. 265-277, Feb 2019, doi: 10.1007/s11255-018-2047-y.
[178] S. Eid *et al.*, "New insights into the mechanisms of diabetic complications: role of lipids and lipid metabolism," *Diabetologia,* vol. 62, no. 9, pp. 1539-1549, 2019.
[179] K. Sharma *et al.*, "Metabolomics reveals signature of mitochondrial dysfunction in diabetic kidney disease," *Journal of the American Society of Nephrology,* vol. 24, no. 11, pp. 1901-1912, 2013.
[180] M. Gillespie *et al.*, "The reactome pathway knowledgebase 2022," *Nucleic Acids Research,* 2021.
[181] P. D. Karp *et al.*, "The BioCyc collection of microbial genomes and metabolic pathways," *Briefings in Bioinformatics,* vol. 20, no. 4, pp. 1085-1093, 2019.

[182] P. Romero, J. Wagg, M. L. Green, D. Kaiser, M. Krummenacker, and P. D. Karp, "Computational prediction of human metabolic pathways from the complete human genome," *Genome Biology,* vol. 6, no. 1, pp. 1-17, 2005.
[183] R. Caspi *et al.*, "The MetaCyc database of metabolic pathways and enzymes-a 2019 update," *Nucleic Acids Research,* vol. 48, no. D1, pp. D445-D453, 2020.
[184] I. M. Keseler *et al.*, "The EcoCyc database in 2021," *Frontiers in Microbiology,* p. 2098, 2021.
[185] J. Pu *et al.*, "MENDA: a comprehensive curated resource of metabolic characterization in depression," *Briefings in Bioinformatics,* vol. 21, no. 4, pp. 1455-1464, 2020.
[186] C. J. Norsigian *et al.*, "BiGG Models 2020: multi-strain genome-scale models and expansion across the phylogenetic tree," *Nucleic Acids Research,* vol. 48, no. D1, pp. D402-D406, 2020.
[187] A. Chang *et al.*, "BRENDA, the ELIXIR core data resource in 2021: new developments and updates," *Nucleic Acids Research,* vol. 49, no. D1, pp. D498-D508, 2021.
[188] H. E. Pence and A. Williams, "ChemSpider: an online chemical information resource," *Journal of Chemical Education*, vol. 87, pp. 1123-1124, 2010.
[189] K. Haug *et al.*, "MetaboLights: a resource evolving in response to the needs of its scientific community," *Nucleic Acids Research,* vol. 48, no. D1, pp. D440-D444, 2020.
[190] M. Sud *et al.*, "Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools," *Nucleic Acids Research,* vol. 44, no. D1, pp. D463-D470, 2016.
[191] L. Cheng *et al.*, "MetSigDis: a manually curated resource for the metabolic signatures of diseases," *Briefings in Bioinformatics,* vol. 20, no. 1, pp. 203-209, 2019.
[192] A. Noronha *et al.*, "The Virtual Metabolic Human database: integrating human and gut microbiome metabolism with nutrition and disease," *Nucleic Acids Research,* vol. 47, no. D1, pp. D614-D624, 2019.
[193] D. N. Slenter *et al.*, "WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research," *Nucleic Acids Research,* vol. 46, no. D1, pp. D661-D667, 2018.
[194] B. Zhang, S. Hu, E. Baskin, A. Patt, J. K. Siddiqui, and E. A. Mathé, "RaMP: a comprehensive relational database of metabolomics pathways for pathway enrichment analysis of genes and metabolites," *Metabolites,* vol. 8, no. 1, p. 16, 2018.
[195] I. Rodchenkov *et al.*, "Pathway Commons 2019 Update: integration, analysis and exploration of pathway data," *Nucleic Acids Research,* vol. 48, no. D1, pp. D489-D497, 2020.
[196] J. Kopka *et al.*, "GMD@ CSB. DB: the Golm metabolome database," *Bioinformatics,* vol. 21, no. 8, pp. 1635-1638, 2005.
[197] A. Mikaia *et al.*, "NIST standard reference database 1A," *Standard Reference Data, NIST, Gaithersburg, MD, USA https://www. nist. gov/srd/nist-standard-referencedatabase-1a,* 2014.
[198] W. Zhou, Y. Ying, and L. Xie, "Spectral database systems: a review," *Applied Spectroscopy Reviews,* vol. 47, no. 8, pp. 654-670, 2012.
[199] G. D. Bader, M. P. Cary, and C. Sander, "Pathguide: a pathway resource list," *Nucleic Acids Research,* vol. 34, no. suppl_1, pp. D504-D506, 2006.

[200] A. Kumar, P. F. Suthers, and C. D. Maranas, "MetRxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases," *BMC Bioinformatics,* vol. 13, no. 1, pp. 1-13, 2012.
[201] R. Alcántara *et al.*, "Rhea—a manually curated resource of biochemical reactions," *Nucleic Acids Research,* vol. 40, no. D1, pp. D754-D760, 2012.
[202] E. Fahy and S. Subramaniam, "RefMet: a reference nomenclature for metabolomics," *Nature Methods,* vol. 17, no. 12, pp. 1173-1174, 2020.
[203] C. Wieder *et al.*, "Pathway analysis in metabolomics: Recommendations for the use of over-representation analysis," *PLoS computational Biology,* vol. 17, no. 9, p. e1009105, 2021.
[204] R. C. Team, "R: A language and environment for statistical computing; 2018," *R Foundation for Statistical Computing: Vienna, Austria*, 2018.
[205] D. K. Barupal, S. Fan, and O. Fiehn, "Integrating bioinformatics approaches for a comprehensive interpretation of metabolomics datasets," *Current Opinion in Biotechnology,* vol. 54, pp. 1-9, 2018.
[206] A. Kaever *et al.*, "MarVis-Pathway: integrative and exploratory pathway analysis of non-targeted metabolomics data," *Metabolomics,* vol. 11, no. 3, pp. 764-777, 2015.
[207] L. Cottret *et al.*, "MetExplore: collaborative edition and exploration of metabolic networks," *Nucleic Acids Research,* vol. 46, no. W1, pp. W495-W502, 2018, doi: 10.1093/nar/gky301.
[208] R. B. M. Aggio, K. Ruggiero, and S. G. Villas-Bôas, "Pathway Activity Profiling (PAPi): from the metabolite profile to the metabolic pathway activity," *Bioinformatics,* vol. 26, no. 23, pp. 2969-2976, 2010.
[209] J. López-Ibáñez, F. Pazos, and M. Chagoyen, "MBROLE 2.0—functional enrichment of chemical compounds," *Nucleic Acids Research,* vol. 44, no. W1, pp. W201-W204, 2016.
[210] N. Kessler *et al.*, "MeltDB 2.0–advances of the metabolomics software system," *Bioinformatics,* vol. 29, no. 19, pp. 2452-2459, 2013.
[211] Z. Pang *et al.*, "MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights," *Nucleic Acids Research,* vol. 49, pp. W388–W396 , 2021.
[212] M. Kankainen, P. Gopalacharyulu, L. Holm, and M. Orešič, "MPEA—metabolite pathway enrichment analysis," *Bioinformatics,* vol. 27, no. 13, pp. 1878-1879, 2011.
[213] G. Kastenmüller, W. Römisch-Margl, B. Wägele, E. Altmaier, and K. Suhre, "metaP-server: a web-based metabolomics data analysis tool," *Journal of Biomedicine and Biotechnology,* vol. 2011, 2011.
[214] B. Wägele, M. Witting, P. Schmitt-Kopplin, and K. Suhre, "MassTRIX reloaded: combined analysis and visualization of transcriptome and metabolome data," *PLoS One,* vol. 7, no. 7, p. e39860, 2012.
[215] D. P. Leader, K. Burgess, D. Creek, and M. P. Barrett, "Pathos: a web facility that uses metabolic maps to display experimental changes in metabolites identified by mass spectrometry," *Rapid Communications in Mass Spectrometry,* vol. 25, no. 22, pp. 3422-3426, 2011.
[216] R. Hernández-de-Diego *et al.*, "PaintOmics 3: a web resource for the pathway analysis and visualization of multi-omics data," *Nucleic Acids Research,* vol. 46, no. W1, pp. W503-W509, 2018.
[217] A. Kamburov, R. Cavill, T. M. D. Ebbels, R. Herwig, and H. C. Keun, "Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA," *Bioinformatics,* vol. 27, no. 20, pp. 2917-2918, 2011.

[218] D. Grapov, K. Wanichthanarak, and O. Fiehn, "MetaMapR: pathway independent metabolomic network analysis incorporating unknowns," *Bioinformatics,* vol. 31, no. 16, pp. 2757-2760, 2015.
[219] V. Danna *et al.*, "leapR: An R Package for Multiomic Pathway Analysis," *Journal of Proteome Research,* vol. 20, no. 4, pp. 2116-2121, 2021.
[220] V. Palombo, M. Milanesi, G. Sferra, S. Capomaccio, S. Sgorlon, and M. D'Andrea, "PANEV: an R package for a pathway-based network visualization," *BMC Bioinformatics,* vol. 21, no. 1, pp. 1-7, 2020.
[221] E. Ulgen, O. Ozisik, and O. U. Sezerman, "pathfindR: An R package for comprehensive identification of enriched pathways in omics data through active subnetworks," *Frontiers in Genetics,* vol. 10, p. 858, 2019.
[222] A. Krämer, J. Green, J. Pollard Jr, and S. Tugendreich, "Causal analysis approaches in ingenuity pathway analysis," *Bioinformatics,* vol. 30, no. 4, pp. 523-530, 2014.
[223] Y. Darzi, I. Letunic, P. Bork, and T. Yamada, "iPath3. 0: interactive pathways explorer v3," *Nucleic Acids Research,* vol. 46, no. W1, pp. W510-W513, 2018.
[224] G. Yu and Q.-Y. He, "ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization," *Molecular BioSystems,* vol. 12, no. 2, pp. 477-479, 2016.
[225] M. Chazalviel *et al.*, "MetExploreViz: web component for interactive metabolic network visualization," *Bioinformatics,* vol. 34, no. 2, pp. 312-313, 2018.
[226] E. Brunk *et al.*, "Recon3D enables a three-dimensional view of gene variation in human metabolism," *Nature Biotechnology,* vol. 36, no. 3, pp. 272-281, 2018.
[227] D. K. Barupal and O. Fiehn, "Chemical Similarity Enrichment Analysis (ChemRICH) as alternative to biochemical pathway mapping for metabolomic datasets," *Scientific Reports,* vol. 7, no. 1, pp. 1-11, 2017.
[228] D. Tenenbaum, S. Runit, M. B. P. Maintainer, M. Carlson, and K. ThirdPartyClient, "Package 'KEGGREST'," *R Foundation for Statistical Computing: Vienna, Austria,* 2019.
[229] B. Wen, Z. Mei, C. Zeng, and S. Liu, "metaX: a flexible and comprehensive software for processing metabolomics data," *BMC Bioinformatics,* vol. 18, no. 1, pp. 1-14, 2017.
[230] M. Leclercq *et al.*, "Large-scale automatic feature selection for biomarker discovery in high-dimensional OMICs data," *Frontiers in Genetics,* vol. 10, p. 452, 2019.
[231] G. Lefort *et al.*, "ASICS: an R package for a whole analysis workflow of 1D 1H NMR spectra," *Bioinformatics,* vol. 35, no. 21, pp. 4356-4363, 2019.
[232] T.-C. Kuo, T.-F. Tian, and Y. J. Tseng, "3Omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data," *BMC Systems Biology,* vol. 7, no. 1, pp. 1-15, 2013.
[233] B. B. Misra, "New software tools, databases, and resources in metabolomics: updates from 2020," *Metabolomics,* vol. 17, no. 5, pp. 1-24, 2021.
[234] S. C. Booth, A. M. Weljie, and R. J. Turner, "Computational tools for the secondary analysis of metabolomics experiments," *Computational and Structural Biotechnology Journal,* vol. 4, no. 5, p. e201301003, 2013.
[235] A.-H. M. Emwas, "The strengths and weaknesses of NMR spectroscopy and mass spectrometry with particular focus on metabolomics research," in *Metabonomics*: Springer, 2015, pp. 161-193.

[236] F. Bhinderwala, N. Wase, C. DiRusso, and R. Powers, "Combining mass spectrometry and NMR improves metabolite detection and annotation," *Journal of Proteome Research,* vol. 17, no. 11, pp. 4017-4022, 2018.
[237] R. M. Gathungu, R. Kautz, B. S. Kristal, S. S. Bird, and P. Vouros, "The integration of LC-MS and NMR for the analysis of low molecular weight trace analytes in complex matrices," *Mass Spectrometry Reviews,* vol. 39, no. 1-2, pp. 35-54, 2020.
[238] A. F. Nassar, T. Wu, S. F. Nassar, and A. V. Wisnewski, "UPLC–MS for metabolomics: a giant step forward in support of pharmaceutical research," *Drug Discovery Today,* vol. 22, no. 2, pp. 463-470, 2017.
[239] E. W. Deutsch, "Mass spectrometer output file format mzML," (in eng), *Methods Mol Biol,* vol. 604, pp. 319-31, 2010, doi: 10.1007/978-1-60761-444-9_22.
[240] P. G. Pedrioli *et al.*, "A common open representation of mass spectrometry data and its application to proteomics research," (in eng), *Nat Biotechnol,* vol. 22, no. 11, pp. 1459-66, Nov 2004, doi: 10.1038/nbt1031.
[241] R. K. Julian *et al.*, "mzData: A standard for the interchange of functional genomics mass spectrometry data," *This issue Nat. Biotechnol,* 2006.
[242] R. Rew and G. Davis, "NetCDF: an interface for scientific data access," *IEEE Computer Graphics and Applications,* vol. 10, no. 4, pp. 76-82, 1990, doi: 10.1109/38.56302.
[243] J. J. Goeman, S. A. Van De Geer, F. De Kort, and H. C. Van Houwelingen, "A global test for groups of genes: testing association with a clinical outcome," *Bioinformatics,* vol. 20, no. 1, pp. 93-99, 2004.
[244] P. D. Karp, P. E. Midford, R. Caspi, and A. Khodursky, "Pathway size matters: the influence of pathway granularity on over-representation (enrichment analysis) statistics," *BMC Genomics,* vol. 22, no. 1, pp. 1-11, 2021.
[245] A. Van Eeckhaut, K. Lanckmans, S. Sarre, I. Smolders, and Y. Michotte, "Validation of bioanalytical LC–MS/MS assays: evaluation of matrix effects," *Journal of Chromatography B,* vol. 877, no. 23, pp. 2198-2207, 2009.
[246] M. Bergman, "The early diabetes intervention program–Is early actually late?," *Diabetes/Metabolism Research and Reviews,* vol. 30, no. 8, pp. 654-658, 2014.
[247] L. L. Gathercole and P. M. Stewart, "Targeting the pre-receptor metabolism of cortisol as a novel therapy in obesity and diabetes," *The Journal of Steroid Biochemistry and Molecular Biology,* vol. 122, no. 1, pp. 21-27, 2010/10/01/ 2010, doi: https://doi.org/10.1016/j.jsbmb.2010.03.060.
[248] R. Pivonello, A. M. Isidori, M. C. De Martino, J. Newell-Price, B. M. K. Biller, and A. Colao, "Complications of Cushing's syndrome: state of the art," *The lancet Diabetes & Endocrinology,* vol. 4, no. 7, pp. 611-629, 2016.
[249] I. Chiodini and V. Morelli, "Subclinical hypercortisolism: how to deal with it?," *Cortisol Excess and Insufficiency,* vol. 46, pp. 28-38, 2016.
[250] R. A. Hackett, M. Kivimäki, M. Kumari, and A. Steptoe, "Diurnal cortisol patterns, future diabetes, and impaired glucose metabolism in the Whitehall II cohort study," *The Journal of Clinical Endocrinology & Metabolism,* vol. 101, no. 2, pp. 619-625, 2016.
[251] B. Zhu *et al.*, "Reduced glycodeoxycholic acid levels are associated with negative clinical outcomes of gestational diabetes mellitus," (in eng), *Journal of Zhejiang University. Science. B,* vol. 22, no. 3, pp. 223-232, 2021, doi: 10.1631/jzus.B2000483.

[252] H. Shapiro, A. A. Kolodziejczyk, D. Halstuch, and E. Elinav, "Bile acids in glucose metabolism in health and disease," *Journal of Experimental Medicine,* vol. 215, no. 2, pp. 383-396, 2018.
[253] W. Zhu *et al.*, "Serum total bile acids associate with risk of incident type 2 diabetes and longitudinal changes in glucose-related metabolic traits," *Journal of Diabetes,* vol. 12, no. 8, pp. 616-625, 2020.
[254] Y. Wu, A. Zhou, L. Tang, Y. Lei, B. Tang, and L. Zhang, "Bile acids: key regulators and novel treatment targets for type 2 diabetes," *Journal of Diabetes Research,* vol. 2020, 2020.
[255] A. Mantovani *et al.*, "Plasma bile acid profile in patients with and without type 2 diabetes," *Metabolites,* vol. 11, no. 7, p. 453, 2021.
[256] N. Le Floc'h, W. Otten, and E. Merlot, "Tryptophan metabolism, from nutrition to potential therapeutic applications," *Amino Acids,* vol. 41, no. 5, p. 1195, 2011.
[257] T. Chen *et al.*, "Tryptophan predicts the risk for future type 2 diabetes," *PloS One,* vol. 11, no. 9, p. e0162192, 2016.
[258] K. Matsuoka *et al.*, "Concentrations of various tryptophan metabolites are higher in patients with diabetes mellitus than in healthy aged male adults," *Diabetology International,* vol. 8, no. 1, pp. 69-75, 2017.
[259] A. Takada, F. Shimizu, J. Masuda, and K. Matsuoka, "Chapter 17 - Plasma Levels of Tryptophan Metabolites in Patients of Type 2 Diabetes Mellitus," in *Bioactive Food as Dietary Interventions for Diabetes (Second Edition)*, R. R. Watson and V. R. Preedy Eds.: Academic Press, pp. 265-276, 2019.
[260] Y. G. C. Varadaiah, S. Sivanesan, S. B. Nayak, and K. R. Thirumalarao, "Purine metabolites can indicate diabetes progression," *Archives of Physiology and Biochemistry,* pp. 1-5, 2019.
[261] F. Ottosson, E. Smith, W. Gallo, C. Fernandez, and O. Melander, "Purine Metabolites and Carnitine Biosynthesis Intermediates Are Biomarkers for Incident Type 2 Diabetes," *The Journal of Clinical Endocrinology & Metabolism,* vol. 104, no. 10, pp. 4921-4930, 2019, doi: 10.1210/jc.2019-00822.
[262] C. Papandreou *et al.*, "Metabolites related to purine catabolism and risk of type 2 diabetes incidence; modifying effects of the TCF7L2-rs7903146 polymorphism," *Scientific Reports,* vol. 9, no. 1, pp. 1-11, 2019.
[263] B. Biondi, G. J. Kahaly, and R. P. Robertson, "Thyroid Dysfunction and Diabetes Mellitus: Two Closely Associated Disorders," (in eng), *Endocrine Reviews,* vol. 40, no. 3, pp. 789-824, 2019, doi: 10.1210/er.2018-00163.
[264] M. Hage, M. S. Zantout, and S. T. Azar, "Thyroid disorders and diabetes mellitus," *Journal of Thyroid Research,* vol. 2011, 2011.
[265] L. Chaker *et al.*, "Thyroid function and risk of type 2 diabetes: a population-based prospective cohort study," *BMC Medicine,* vol. 14, no. 1, pp. 1-8, 2016.
[266] L. Wen and F. S. Wong, "Dietary short-chain fatty acids protect against type 1 diabetes," *Nature Immunology,* vol. 18, no. 5, pp. 484-486, 2017.
[267] Q. Yang, J. Ouyang, F. Sun, and J. Yang, "Short-Chain Fatty Acids: A Soldier Fighting Against Inflammation and Protecting From Tumorigenesis in People With Diabetes," *Frontiers in Immunology,* vol. 11, p.3139, 2020.
[268] H. Zhou *et al.*, "Short-chain fatty acids can improve lipid and glucose metabolism independently of the pig gut microbiota," *Journal of Animal Science and Biotechnology,* vol. 12, no. 1, pp. 1-14, 2021.

[269] D. Salamone, A. A. Rivellese, and C. Vetrani, "The relationship between gut microbiota, short-chain fatty acids and type 2 diabetes mellitus: the possible role of dietary fibre," *Acta Diabetologica,* pp. 1-8, 2021.
[270] B. A. Menge *et al.*, "Selective amino acid deficiency in patients with impaired glucose tolerance and type 2 diabetes," *Regulatory Peptides,* vol. 160, no. 1, pp. 75-80, 2010/02/25/ 2010, doi: https://doi.org/10.1016/j.regpep.2009.08.001.
[271] H. Chilukuri, M. J. Kulkarni, and M. Fernandes, "Revisiting amino acids and peptides as anti-glycation agents," (in eng), *MedChemComm,* vol. 9, no. 4, pp. 614-624, 2018, doi: 10.1039/c7md00514h.
[272] S. C. Connor, M. K. Hansen, A. Corner, R. F. Smith, and T. E. Ryan, "Integration of metabolomics and transcriptomics data to aid biomarker discovery in type 2 diabetes," *Molecular BioSystems,* vol. 6, no. 5, pp. 909-921, 2010.
[273] S. Park, K. C. Sadanala, and E.-K. Kim, "A Metabolomic Approach to Understanding the Metabolic Link between Obesity and Diabetes," (in eng), *Molecules and Cells,* vol. 38, no. 7, pp. 587-596, 2015, doi: 10.14348/molcells.2015.0126.
[274] M. A. Gentile *et al.*, "Androgen-mediated improvement of body composition and muscle function involves a novel early transcriptional program including IGF1, mechano growth factor, and induction of β-catenin," *Journal of Molecular Endocrinology,* vol. 44, no. 1, p. 55, 2010.
[275] T. Sathyapalan, E. H. Dickerson, S. M. Maguiness, J. Robinson, Y. H. Z. Dakroury, and S. L. Atkin, "Androstenedione and testosterone levels correlate with in vitro fertilization rates in insulin-resistant women," *BMJ Open Diabetes Research and Care,* vol. 5, no. 1, p. e000387, 2017.
[276] P. K. Ganguly, K. S. Dhalla, I. R. Innes, R. E. Beamish, and N. S. Dhalla, "Altered norepinephrine turnover and metabolism in diabetic cardiomyopathy," *Circulation Research,* vol. 59, no. 6, pp. 684-693, 1986.
[277] K. Wang *et al.*, "Differences in proximal tubular solute clearance across common etiologies of chronic kidney disease," *Nephrology Dialysis Transplantation,* vol. 35, no. 11, pp. 1916-1923, 2020.
[278] W. Al-Badr and K. J. Martin, "Vitamin D and kidney disease," *Clinical Journal of the American Society of Nephrology,* vol. 3, no. 5, pp. 1555-1560, 2008.
[279] N. R. Dash and M. T. Al Bataineh, "Metagenomic Analysis of the Gut Microbiome Reveals Enrichment of Menaquinones (Vitamin K2) Pathway in Diabetes Mellitus," *Diabetes & Metabolism Journal,* vol. 45, no. 1, pp. 77-85, 2021.
[280] B. Hocher and J. Adamski, "Metabolomics for clinical use and research in chronic kidney disease," *Nature Reviews Nephrology,* vol. 13, no. 5, pp. 269-284, 2017.
[281] S. Yang *et al.*, "Mitochondria: a novel therapeutic target in diabetic nephropathy," *Current Medicinal Chemistry,* vol. 24, no. 29, pp. 3185-3202, 2017.
[282] J. M. Forbes and D. R. Thorburn, "Mitochondrial dysfunction in diabetic kidney disease," *Nature Reviews Nephrology,* vol. 14, no. 5, pp. 291-312, 2018/05/01 2018, doi: 10.1038/nrneph.2018.9.
[283] P. Z. Wei and C. C. Szeto, "Mitochondrial dysfunction in diabetic kidney disease," *Clinica Chimica Acta,* vol. 496, pp. 108-116, 2019.

[284] L. Li *et al.*, "Metabolomics reveal mitochondrial and fatty acid metabolism disorders that contribute to the development of DKD in T2DM patients," *Molecular BioSystems,* vol. 13, no. 11, pp. 2392-2400, 2017.
[285] A. L. T. Dos Santos *et al.*, "Low linolenic and linoleic acid consumption are associated with chronic kidney disease in patients with type 2 diabetes," *PLoS One,* vol. 13, no. 8, p. e0195249, 2018.
[286] Q. Sha, J. Lyu, M. Zhao, H. Li, M. Guo, and Q. Sun, "Multi-Omics Analysis of Diabetic Nephropathy Reveals Potential New Mechanisms and Drug Targets," *Frontiers in Genetics,* vol. 11, p. 1605, 2020.
[287] C. H. Johnson, A. D. Patterson, J. R. Idle, and F. J. Gonzalez, "Xenobiotic metabolomics: major impact on the metabolome," *Annual Review of Pharmacology and Toxicology,* vol. 52, pp. 37-56, 2012.
[288] C. H. Johnson, J. Ivanisevic, and G. Siuzdak, "Metabolomics: beyond biomarkers and towards mechanisms," *Nature Reviews Molecular Cell Biology,* vol. 17, no. 7, pp. 451-459, 2016.
[289] P. Vineis, O. Robinson, M. Chadeau-Hyam, A. Dehghan, I. Mudway, and S. Dagnino, "What is new in the exposome?," *Environment International,* vol. 143, p. 105887, 2020.
[290] D. I. Walker, D. Valvi, N. Rothman, Q. Lan, G. W. Miller, and D. P. Jones, "The metabolome: A key measure for exposome research in epidemiology," *Current Epidemiology Reports,* vol. 6, no. 2, pp. 93-103, 2019.
[291] M. D. Stobbe, S. M. Houten, G. A. Jansen, A. H. C. van Kampen, and P. D. Moerland, "Critical assessment of human metabolic pathway databases: a stepping stone for future integration," *BMC Systems Biology,* vol. 5, no. 1, pp. 1-19, 2011.
[292] N. Pham, R. G. A. van Heck, J. C. J. van Dam, P. J. Schaap, E. Saccenti, and M. Suarez-Diez, "Consistency, inconsistency, and ambiguity of metabolite names in biochemical databases used for genome-scale metabolic modelling," *Metabolites,* vol. 9, no. 2, p. 28, 2019.
[293] A. Kamburov, C. Wierling, H. Lehrach, and R. Herwig, "ConsensusPathDB—a database for integrating human functional interaction networks," *Nucleic Acids Research,* vol. 37, no. suppl_1, pp. D623-D628, 2009.
[294] D. Domingo-Fernández, S. Mubeen, J. Marín-Llaó, C. T. Hoyt, and M. Hofmann-Apitius, "PathMe: merging and exploring mechanistic pathway knowledge," *BMC Bioinformatics,* vol. 20, no. 1, pp. 1-12, 2019.
[295] K. Wanichthanarak, S. Fan, D. Grapov, D. K. Barupal, and O. Fiehn, "Metabox: A toolbox for metabolomic data analysis, interpretation and integrative exploration," *PloS One,* vol. 12, no. 1, p. e0171046, 2017.
[296] M. Kutmon *et al.*, "PathVisio 3: an extendable pathway analysis toolbox," *PLoS Computational Biology,* vol. 11, no. 2, p. e1004085, 2015.
[297] G. B. Kim, W. J. Kim, H. U. Kim, and S. Y. Lee, "Machine learning applications in systems metabolic engineering," *Current Opinion in Biotechnology,* vol. 64, pp. 1-9, 2020/08/01/ 2020, doi: https://doi.org/10.1016/j.copbio.2019.08.010.
[298] C. E. Lawson *et al.*, "Machine learning for metabolic engineering: A review," *Metabolic Engineering,* vol. 63, pp. 34-60, 2021.
[299] H. B. Burke, "Predicting clinical outcomes using molecular biomarkers," *Biomarkers in Cancer,* vol. 8, pp. BIC-S33380, 2016.
[300] M. J. Selleck, M. Senthil, and N. R. Wall, "Making meaningful clinical use of biomarkers," *Biomarker Insights,* vol. 12, p. 1177271917715236, 2017.

Vita

Bayan Hassan Banimfreg was born in 1987 in Irbid, Jordan. She received her primary and secondary education in Irbid, Jordan. She received her B.Sc. degree in Electrical Engineering from Al Yarmouk University in 2010. She also received her MSc degree in Electrical Engineering from the University of Texas at San Antonio (UTSA) in 2014. In addition, she received a Master of Business Administration (MBA) from Webster University in 2016.

In August 2018, she joined the Engineering Systems Management Ph.D. program at the American University of Sharjah as a graduate teaching and research assistant. During her Ph.D. study, she co-authored three papers. Her research interests are in big data analytics in healthcare.