Improving a biosignature for respiratory chain deficiencies in a South African cohort

Jaundrie Fourie

22262385

Dissertation submitted in partial fulfilment of the requirements for the degree *Magister Scientiae* in Biochemistry at the Potchefstroom Campus of the North-West University

Supervisor: Dr R Louw

Co-supervisor: Dr Z Lindeque

Assistant supervisor: Prof I Smuts

MEI 2016

ABSTRACT

Mitochondria are the cell's main energy producing site, found in the cytoplasm of nearly all eukaryotic cells. These organelles generate cellular energy in the form of adenosine triphosphate (ATP), mostly by means of the oxidative phosphorylation (OXPHOS) system, consisting of the respiratory chain (RC) and ATP synthase complex. When one of the four complexes (which form the RC) becomes impaired, it is called a respiratory chain deficiency (RCD). Diagnosing RCDs is a major challenge and requires a multi-disciplinary approach, which includes clinical, histochemical, molecular and biochemical assessment. The golden standard for diagnosing a RCD is enzyme analyses on a muscle sample obtained from a muscle biopsy, which is an invasive procedure. A urinary biosignature was proposed, that has the potential to be used as a screening tool for selecting patients that have a potential RCD. The proposed biosignature however consisted of 12 features that still required verification. The aim of this study was to improve this proposed biosignature, by verifying the 12 features and expanding the biosignature by analysing the identical sample cohort with alternative analytical platforms in order to discover additional features. This study was conducted in two separate phases, a verification phase and an expanding phase.

Two levels of verification were performed for the LC-MS features of the proposed biosignature. Five of the 12 features could be verified and only one could be identified to a certain extent. For the expansion part of this study, two additional platforms were used, a gas-chromatography mass spectrometry (GC-MS) and nuclear magnetic resonance (NMR) analysis. Following evaluation of the methods used and sample analysis, a number of data mining steps and statistical analyses were performed to compile a list of top ranked features for each platform. The LC-MS, GC-MS and NMR features were considered for the improved biosignature, by using a variety of statistical analyses the number and best combination of features were selected. A list of five features was compiled and the discriminative power was evaluated. Results indicated that the improved biosignature was unable to classify samples 100% accurately with some of the clinical control samples classified as RCDs, however the biosignature could still be helpful in limiting the inclusion of CRC patients in the biopsy process which gives it the potential to be used in the diagnostic workflow.

Key words: respiratory chain deficiency; metabolomics; urinary biosignature; LC-MS, GC-MS, NMR.

“*The purpose of life is not to be happy. It is to be useful, to be honorable, to be compassionate, to have it make some difference that you have lived and lived well*.”

— Ralph Waldo Emerson

ACKNOWLEDGEMENTS

First and foremost I would like to thank our Heavenly Father, for the talents He provided me with. For giving me the strength to complete this part of my studies. For all the wonderful people He put into my life and for always carrying me.

I would like to thank the following people, without whom this dissertation would not have been possible:

To my supervisor, Dr Roan Louw, thank you for your patience, advice and guidance over the past years. Also thank you for your support and encouragement when I needed it the most.

Dr Zander Lindeque, my co-supervisor, thank you for your guidance, advice and motivation during these past two years, it's been an enormous learning experience and I'm grateful for everything.

Mr Peet Jansen Van Rensburg, thank you for your patience, help and all the contributions you made, it is deeply appreciated.

Mr Shayne Mason, thank you for your assistance with the NMR work.

Thank you to Ms. Hayley van Dyk for the language and spelling editing of my dissertation.

To my fellow members of the Mitochondrial Research Laboratory, you supported me every day without even realising it, I am going to miss you dearly.

All my friends and family, near and far, who have supported me in some way, but with special thanks to, My parents: Hans and Mandie, for their unconditional love, support and always believing in me. I will always be grateful for what you have provided me with and will never be able to say thank you enough. I love you.

To my sister Melissa, thank you for all your love and support.

Leonie Venter and Karien Esterhuizen, for being the unofficial supervisors, thank you for your advice, encouragement and friendship. I learned a lot from both of you as individuals and scientists.

Abel, Hayley, Lindi, Taryn and Vida, you started this journey with me. We laughed, cried and complained throughout our post-graduate studies. Guess what, we made it!

TABLE OF CONTENTS

LIST OF TABLES

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

LIST OF FIGURES

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

LIST OF EQUATIONS

Chapter 3

Chapter 4

ABBREVIATIONS

A

ACN	:	Acetonitrile
Acetyl-CoA		Acetyl-coenzyme A
ADP	:	Adenosine diphosphate
ATP	:	Adenosine triphosphate
AUC	:	Area under the curve

C

CoA	:	Coenzyme A
Complex 1	:	CI: NADH coenzyme Q reductase)
Complex 2	:	CII: Succinate-CoQ reductase complex
Complex 3	:	CII: Reduced CoQ cytochrome *c* reductase complex
Complex 4	:	CIV: Cytochrome *c* oxidase
Complex 5	:	CV: ATP synthase complex
CoQ	:	Coenzyme Q
CRC	:	Clinical referred control
CS	:	Citrate synthase
CV	:	Coefficient of variation

D

DNA	:	Deoxyribonucleic acid

E

ESI:	:	Electron spray ionisation

F

FADH	:	Flavin adenine dinucleotide (reduced)

G

GC-MS	:	Gas chromatography mass spectrometer

H

HILIC		Hydrophilic-interaction chromatography
HPLC	:	High-performance liquid chromatography
H_2O	:	Water

I

IS	:	Internal standard
IMM		Inner mitochondrial membrane

L

L	:	Litre
LC:	:	Liquid chromatography
LC-MS	:	Liquid chromatography mass spectrometry
LC-Q-TOF	:	Liquid chromatography quadrupole time-of-flight
LHON	:	Leber's hereditary optic neuropathy
LS	:	Leigh Syndrome

M

MDC	:	Mitochondrial Disease Criteria
MELAS	:	Mitochondrial myopathy, encephalopathy, lactic acidosis, and stroke syndrome
MERRF	:	Mitochondrial encephalomyopathy characterized by ragged red fibers in damaged muscle
METLIN:	:	Metabolite and tandem MS database
mg	:	Milligrams
MilliQ	:	Millipore
MS	:	Mass spectrometry
MS/MS	:	Tandem mass spectrometry
mtDNA	:	Mitochondrial deoxyribonucleic acid
m/z	:	Mass to charge

N

NAD		Nicotinamide adenine dinucleotide (oxidised)
NADH	:	Nicotinamide adenine dinucleotide (reduced)
nDNA	:	Nuclear deoxyribonucleic acid
NMR	:	Nuclear magnetic resonance

O

OMM	:	Outer mitochondrial membrane
OXPHOS	:	Oxidative phosphorylation

P

PC	:	Principal component
PCA	:	Principal component analyses
PCR	:	Polymerase chain reaction
PLS-DA	:	Partial least square discriminant analysis
ppm	:	Parts per million

Q

QC	:	Quality control
Q-TOF	:	Quadrupole time of flight

R

RC	:	Respiratory chain (CI-CIV)
RCD	:	Respiratory chain deficiency
RCDs	:	Respiratory chain deficiencies
RF	:	Random forest
RNA	:	Ribonucleic acid
ROC	:	Receiving operating characteristic
ROCCET	:	ROC curve explorer and tester
RP	:	Reverse phase
RRF	:	Ragged red fibers
RT	:	Retention time

S

SDH	:	Succinate dehydrogenase
SVM	:	Support vector machine

T

TCA	:	Tricarboxylic acid
TOM	:	Translocase of the outer membrane
TOF		Time of flight

V

VIP	:	Variable important in projection

CHAPTER 1

INTRODUCTION

1

The hub of energy metabolism, the mitochondrion, is found in virtually all eukaryotic cells, with the exception being erythrocytes. The mitochondrion generates cellular energy in the form of adenosine triphosphate (ATP), mostly by means of the oxidative phosphorylation (OXPHOS) system that is located in the inner mitochondrial membrane. The respiratory chain (consisting of complexes I, II, III and IV) and ATP synthase (complex V) are collectively known as the OXPHOS system, and are encoded by both nuclear DNA (nDNA) and mitochondrial DNA (mtDNA). The number of mitochondria per cell, ranging from hundreds to thousands, is controlled by the energy requirements of specific tissues, with the greatest abundance of mitochondria found in metabolically active tissue (Pieczenik and Neustadt, 2007). Mitochondrial disease can occur when there is a defect in any of the numerous mitochondrial pathways, caused by spontaneous or inherited mutations. Respiratory chain deficiencies (RCDs) are the largest subgroup of mitochondrial disease and occur when one of the four respiratory chain complexes become impaired. RCDs are considered to be one of the most common forms of inherited metabolic diseases, with an estimated incidence of one in every 5 000 - 10 000 live births (Applegarth *et al.,* 2000; Darin *et al*., 2001; Gorman *et al*., 2015; Schaefer *et al.,* 2008; Skladal *et al*., 2003; Suomalainen, 2011b). RCDs can theoretically present with any symptom, in any organ and at any age (Rötig and Munnich, 2003).

Diagnosis of RCDs has been a major challenge for many years and requires a multi-disciplinary approach. Enzyme analyses in muscle biopsies are generally considered the golden standard for diagnosing RCDs. However, obtaining a muscle biopsy is an invasive procedure with numerous limitations and complications. A urinary biosignature (list of biomarkers) for RCDs was recently reported by Venter *et al*. (2015). This biosignature is far less invasive than obtaining a muscle biopsy and thus has the potential to be used as a screening tool for selecting patients with a potential RCD. Theoretically, patients identified using this screening tool can thereafter be subjected to a muscle biopsy in order to confirm the presence of an RCD. Since the reported biosignature has not been verified or validated, this study was conducted to verify and improve it.

The structure of the thesis is as follows: Chapter 2 consists of a literature review, with a complete outline of the experimental approach, aim and objectives. Chapter 3 contains all methods and results of the verification phase. In Chapters 4 and 5, all the methodological development steps are described along with the results obtained using gas chromatography-mass spectrometry and nuclear magnetic resonance analysis, respectively, to improve the existing biosignature. Chapter 6 is the concluding chapter which summarizes all of the findings of the previously mentioned chapters, followed by future prospects and recommendations.

CHAPTER 2

LITERATURE OVERVIEW

2

2.1 Introduction

Chapter 1 presented the background and research rationale of this study. This chapter will provide a detailed overview of the relevant literature for investigating respiratory chain deficiencies, where focus will be placed on the mitochondrion, OXPHOS system, mitochondrial defects, respiratory chain deficiencies and metabolomics. At the end of this chapter, the problem statement, aim and objectives will be given, followed by the experimental strategy that was followed.

2.2 The Mitochondrion

Mitochondria are small, double membrane organelles found in the cytoplasm of all eukaryotic cells (with the exception of mature erythrocytes) and are the cell's main energy producing site (Cohen and Gold, 2001; Herrmann & Neuper, 2000). Mitochondria provide more than 90% of the required energy of cells by metabolizing nutrients or body reserves into ATP. The number of mitochondria per cell is controlled by the specific tissue's energy requirements, ranging from hundreds to thousands in each cell, with the greatest abundance of mitochondria found in metabolically active tissue, e.g. skeletal muscle, cardiac muscle, the brain and liver (Pieczenik and Nestadt, 2007; Spinazzola and Zeviani 2009).

2.2.1 Structure of mitochondria

The mitochondrion is approximately 0.5-1 µm in diameter and 7 µm in length with a rod or sphere shape. Mitochondria vary in shape and number but have the same basic architecture, as illustrated in Figure 2.1. The mitochondrion is composed of two phospholipid bilayer membranes, namely the outer mitochondrial membrane (OMM) and the inner mitochondrial membrane (IMM), which are involved in the key functions of the mitochondrion and have different biochemical functions (Krauss, 2001). The OMM contains porins that allow molecules smaller than 5 kilo Dalton (kDa) to freely diffuse through, while larger proteins enter by binding to transporters (called translocases) located within the membrane (Bolisetty and James, 2013).

The IMM is more resistant to ions and molecules than the OMM, allowing only water, oxygen and carbon dioxide to freely pass through, therefore playing a central role in compartmentalization. The OMM and IMM subdivide the mitochondrion into two distinct sub-compartments, namely, the mitochondrial matrix and the intermembrane space (Herrmann & Neuper, 2000). The IMM encloses the mitochondrial matrix and is folded inward forming cristae, which increase its surface area. The inner membrane houses the key enzymatic machinery of the OXPHOS system and acts as an electrical insulator as well as a chemical barrier, thus assisting in the maintenance of an electrochemical gradient, which is essential for the generation of ATP (Bolisetty and James, 2013).

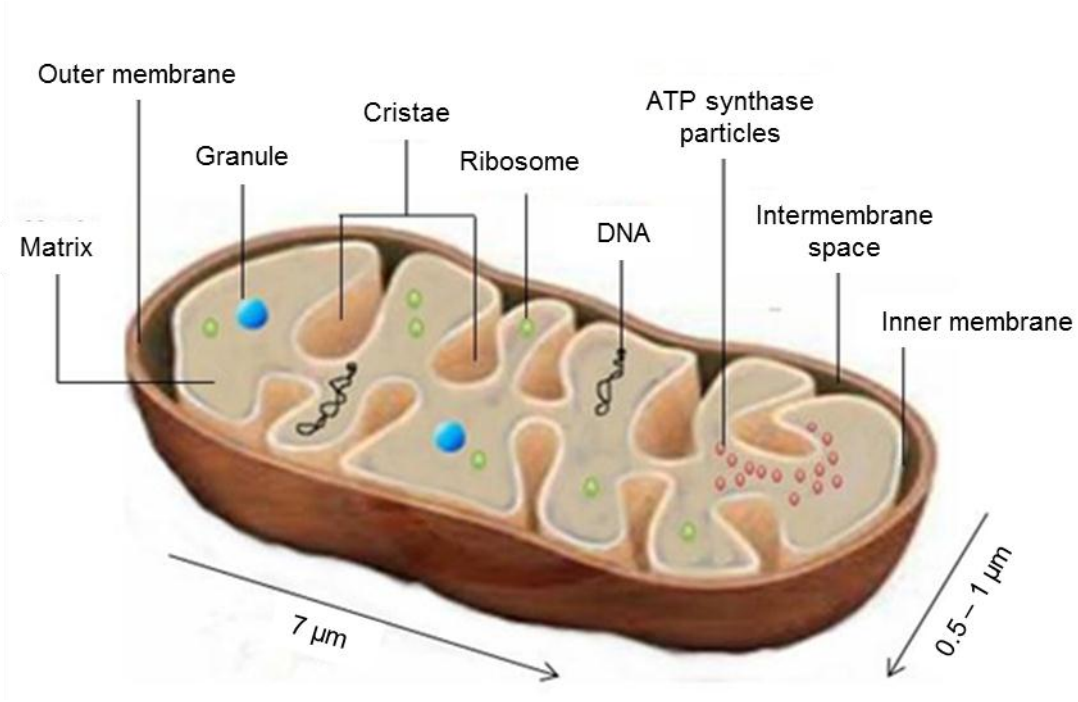

Figure 2.1: The structure of the mitochondrion. The different compartments of the mitochondrion are displayed. These include the OMM, IMM, intermembrane space, mtDNA, ribosomes and cristae. (Adapted from Bandyopachyay and Dutta, 2010)

2.2.2 Function of the mitochondrion

The production of energy in the form of ATP, via the OXPHOS system, is one of the most important functions of the mitochondrion and will be described in the following section (Dimauro and Scon, 2003; Saeed and Dinger, 2013). Mitochondria harbour other processes such as the tricarboxylic acid (TCA) cycle, haeme biosynthesis, degradation of amino acids, pyruvate oxidation, uric acid cycle and metabolism of fatty acids (β-oxidation) and steroids. The mitochondrion also plays a role in the signalling and triggering of programmed cell death where

cytochrome *c* is released into the cytosol and activates intrinsic pathways of apoptosis (Saeed and Singer, 2013). Other functions include the production of certain hormones, biosynthesis of iron-sulphur clusters, production of heat and reactive oxygen species. The generation of heat is achieved by uncoupling of the OXPHOS system which leads to thermogenesis (Bandyopachyay and Dutta, 2010).

2.3 Energy production in mitochondrion

Cellular energy is produced in the form of ATP by three metabolic processes, namely: glycolysis, the mitochondrial TCA cycle and the OXPHOS system. Figure 2.2 illustrates the breakdown of nutrients through these pathways in order to produce the energy required for normal cellular function. Glycolysis takes place in the cytosol and is thus not included in this discussion.

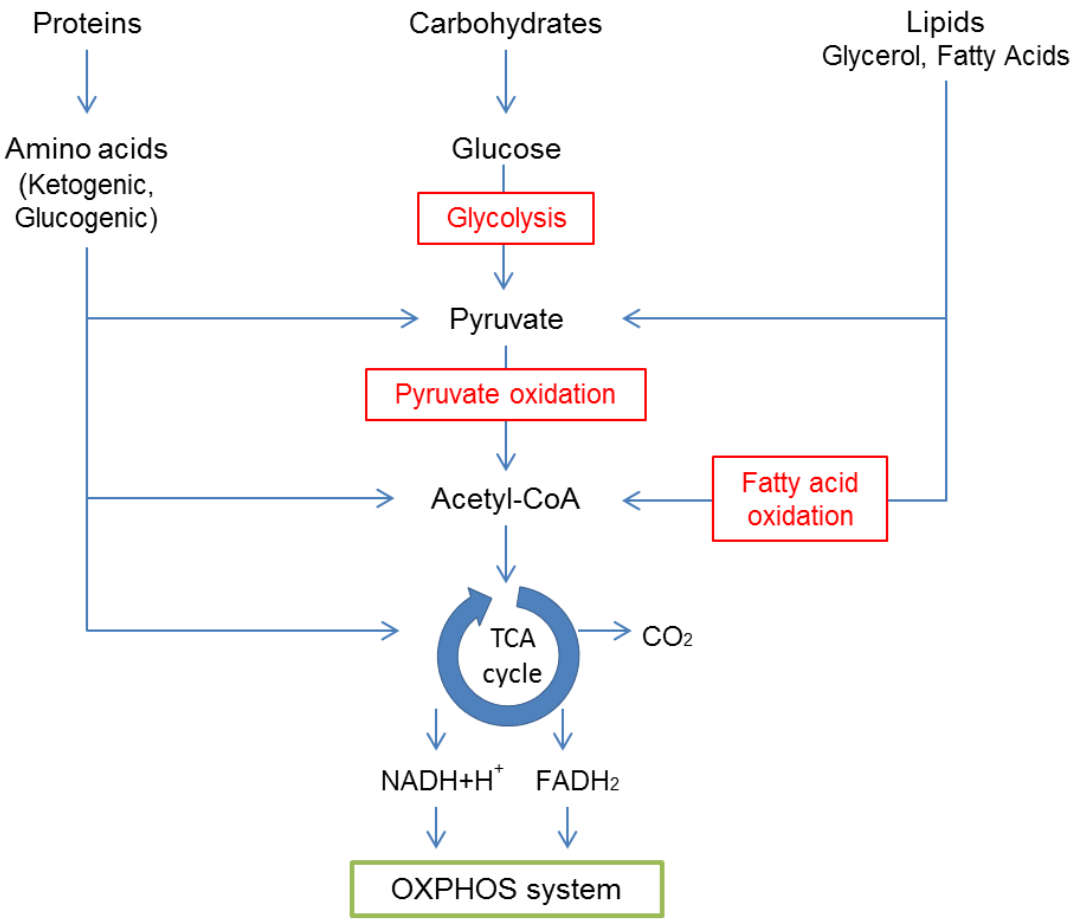

Figure 2.2: Metabolic pathways that are responsible for energy production. This simplified figure of the major ATP-producing pathways illustrates how nutrients ingested are catabolized to form energy. The TCA cycle and OXPHOS system reside in the mitochondria, contrary to glycolysis, which takes place in the cytosol (Adapted from Garret and Grisham, 2008).

2.3.1 Carbohydrates, amino acids and lipids

Nutrients such as proteins, lipids and carbohydrates are metabolized in order to produce energy, as shown in Figure 2.2. Carbohydrates are broken down to glucose which is in turn broken down to pyruvate, via glycolysis. Pyruvate is transported into the mitochondria and then converted to acetyl-coenzyme A (acetyl-CoA) by the enzyme pyruvate dehydrogenase (Fernie *et al*., 2004; McInnes, 2013).

The building blocks of proteins are amino acids. Excess proteins can be used as a source of metabolic energy by removing the α-amino groups of amino acids via the process of deamination and converting the remaining carbon skeletons to compounds such as glucose, pyruvate and TCA cycle intermediates. Ketogenic amino acids are those amino acids that are broken down to fatty acids and ketone bodies, ultimately producing acetyl-CoA. Glucogenic amino acids are the amino acids that are degraded to TCA cycle intermediates, pyruvate and glucose (Berg *et al.*, 2002a).

Lipids are broken down to form glycerol and fatty acids. Glycerol is converted to pyruvate, while fatty acids are broken down by the process of β-oxidation. Fatty acids are transported across the OMM into the intermembrane space by palmitoyl transferase I, where they are transformed into acyl-carnitine. The acyl-carnitine is then transported across the IMM by carnitine acyl translocase, into the mitochondrial matrix, where it is subsequently converted to acyl-CoA by palmitoyl transferase II. Following β-oxidation, acetyl-CoA enters the TCA cycle where the electron carriers nicotinamide adenine dinucleotide (NADH) and flavin adenine dinucleotide ($FADH_2$) are generated (Pieczenik and Nestadt, 2007). Ketone bodies are high energy compounds that are produced in the liver and kidneys as a by-product of β-oxidation and have the ability to serve as an alternative energy source in tissue when glucose oxidation is impaired (McInnes, 2013).

2.3.2 The tricarboxylic acid cycle

The TCA cycle, also referred to as the Krebs-cycle, plays a role in the conversion of lipids and carbohydrates into ATP, but its major function is the production of NADH and $FADH_2$ which are able to enter the electron transport chain, and thus produce large amounts of ATP. The pathway begins with the catabolism of glucose into two molecules of pyruvate in the cytosol. These pyruvate molecules move through the double membrane of the mitochondria where they encounter two enzymes, pyruvate carboxylase and pyruvate dehydrogenase.

When ATP is sufficient, pyruvate carboxylase is activated and pyruvate is shuttled in the direction of gluconeogenesis. However, when there is a high demand for ATP, pyruvate passes through pyruvate dehydrogenase (a complex of enzymes that require multiple coenzymes and substrates to function) to produce two acetyl-CoA molecules (Pieczenik and Neustadt, 2007). Acetyl-CoA enters the TCA cycle by reacting with the four-carbon molecule oxaloacetate, a reaction catalysed by citrate synthase, to form a six-carbon molecule called citric acid. Citric acid is further processed in a series of enzyme-catalysed reactions including several dehydrogenases, as illustrated in Figure 2.3 (Berg *et al.*, 2002a). Each acetyl-CoA molecule produces three molecules of NADH and two molecules of $FADH_2$, thus providing a total of six NADH and four $FADH_2$ molecules per pyruvate molecule (Pieczenik and Neustadt, 2007).

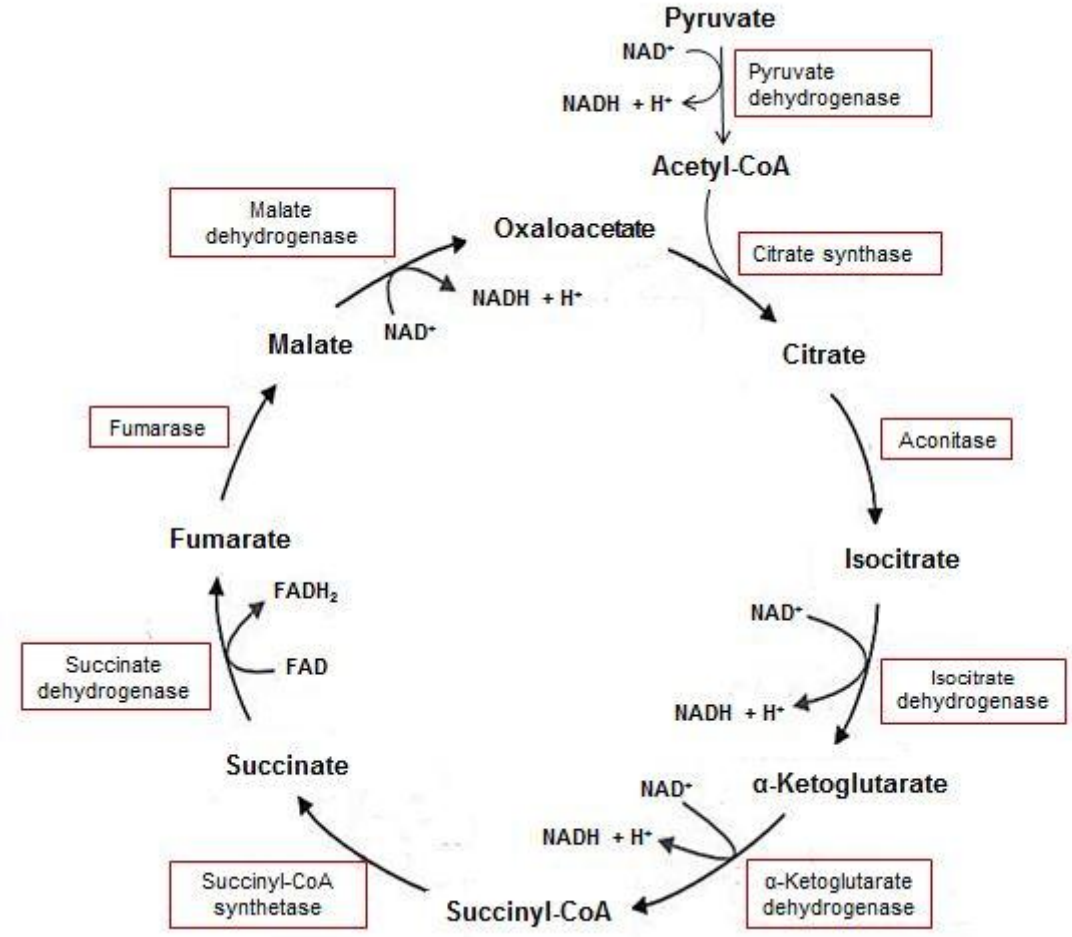

Figure 2.3: Illustration of the TCA cycle in the mitochondrion. The cycle consists of the stepwise breakdown of pyruvate to produce electron carriers capable of entering the electron transport chain (Adapted from Garret and Grisham, 2008).

2.3.3 The OXPHOS system

The OXPHOS system consists of the respiratory chain (RC), and ATP synthase (also known as complex V), as well as the two shuttle molecules cytochrome *c* and coenzyme Q10. The RC (Figure 2.4) consists of four multi-subunit protein complexes, namely, complex I (NADH-ubiquinone oxidoreductase), complex II (succinate:ubiquinone oxidoreductase), complex III (cytochrome bc1) and complex IV (cytochrome oxidase) (Saraste, 1999). Electrons derived from NADH and $FADH_2$ flow through the complexes of the RC, where NADH donates electrons to complex I, and complex II receives electrons from $FADH_2$. Electrons are passed along by ubiquinone (coenzyme Q10) from complexes I and II to complex III, and by cytochrome *c (*an iron containing heme protein) from complex III to complex IV. As electrons are transferred along the RC, protons are pumped across the IMM by complexes I, III and IV, generating an electrochemical gradient. This gradient then provides the energy required to drive ATP synthase and allow movement of protons through complex V into the matrix with the subsequent production of energy during the phosphorylation of adenosine diphosphate (ADP) to ATP (Pieczenik and Neustadt, 2007).

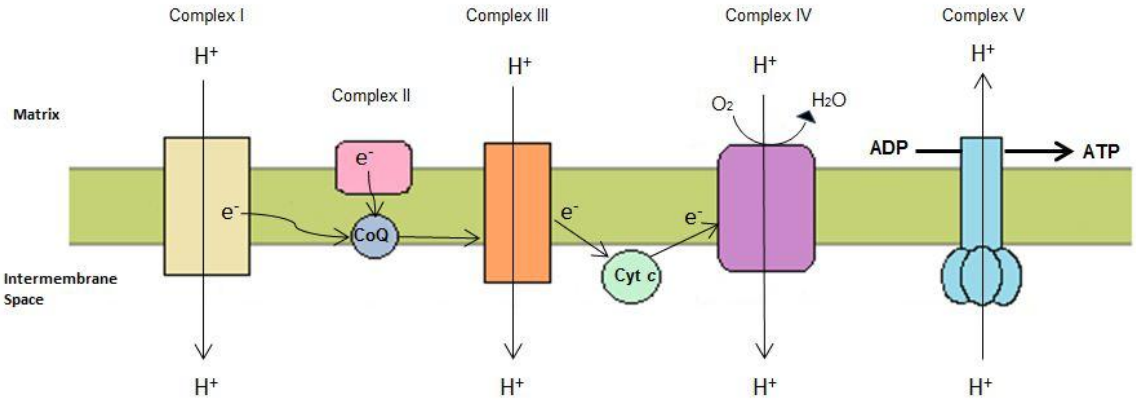

Figure 2.4: The complexes of the OXPHOS system. This figure shows the simplified representation of the OXPHOS system which consists of the four complexes of the RC (Complex I-CIV) and ATP synthase (Complex V). ATP is formed from ADP by complex V. ADP: adenosine diphosphate; ATP: adenosine triphosphate; O_2: oxygen; H_2O: water; CoQ: coenzymeQ; Cyt c: cytochrome c; NADH: reduced nicotinamide adenine dinucleotide; FADH2: flavine adenine dinucleotide; (Adapted from DiMauro 2004).

2.4 Mitochondrial genetics

The human mitochondrial genome is a closed, circular, double-stranded DNA built of 16 569 base pairs. Mitochondrial DNA contains 37 genes that encode 13 proteins essential for the OXPHOS system, 22 transfer RNAs and two ribosomal RNAs. The remaining proteins needed for mitochondrial maintenance and metabolism are encoded by nDNA and are transported into the mitochondria (Greaves *et al*., 2012).

The complexes of the OXPHOS system consist of subunits that are encoded by both nDNA and mtDNA. Complex I consist of 45 subunits, seven of which are encoded by mtDNA, while complex II consists of four subunits, none of which are mtDNA encoded. Complex III consists of 11 subunits, of which only one is mtDNA encoded. Complex IV consists of 13 subunits, three of which are encoded by mtDNA, while complex V consists of approximately 16 subunits, of which two are mtDNA encoded. The remaining subunits that make up each complex are encoded by nDNA (DiMauro and Schon, 2003; Fernández-Vizarra *et al*., 2009).

nDNA is wrapped around histones that shield nDNA from harmful free radicals. In contrast, mtDNA are not protected by histones and are thus more susceptible to damage (Pieczenik and Neustadt, 2007). In addition, the lack of protection and repair mechanisms within mtDNA, combined with mtDNAs close proximity to the ETC (where it is more exposed to free radicals), leads to an mtDNA mutation rate of six to seven times higher than nDNA (Cohen and Gold, 2001). Both nDNA and mtDNA mutations are associated with mitochondrial disease (Taylor and Turnbull, 2005). .

Three major aspects distinguish mitochondrial genetics from Mendelian genetics: maternal inheritance, mitotic segregation and heteroplasmy. The mitochondrial genome is inherited maternally since during fertilization, all mitochondria found in the zygote are obtained from the oocyte. Mutations in mtDNA are therefore transmitted from mother to child and only daughters will then transmit the mutation to their progeny (DiMauro and Schon, 2003). Disease associated with nDNA mutations follow the Mendelian pattern of inheritance (Sue & Sohon, 2000). Mitochondrial disease is therefore inherited maternally, autosomal recessively, autosomal dominantly or in an X-linked manner (Rötig and Munnich, 2003; Suomalainen, 2011b).

When all of the mtDNA in a cell is identical, it is referred to as homoplasmy. Somatic mutations in mtDNA can either be inherited, or they can accumulate with aging. The cell may contain wild-type as well as mutated mtDNA, which in this case is referred to as heteroplasmy (Schapira, 2006). A pathogenic mtDNA mutation has to accumulate past a minimum critical number to

result in mitochondrial dysfunction; this is known as the threshold effect. Organs with high energy demands, such as skeletal muscle, heart, brain, retina, renal tubes and endocrine glands have a lower threshold for disease (DiMauro, 2006). When the amount of mutant mtDNA increases, the mitochondrial function decreases and the threshold is exceeded, resulting in possible initiation of necrosis or apoptosis (Wallace and Fan, 2010). Another aspect of mitochondrial genetics is mitotic segregation that takes place during cell division. When cells divide, the mitochondria are redistributed randomly amongst daughter cells. During cell replication in heteroplasmic cells, the mtDNA genotype can shift, resulting in some lineages remaining heteroplasmic and others drifting towards wild-type mtDNA and becoming homoplasmic (Rötig and Munnich, 2003).

2.5 Mitochondrial disease

According to Liang *et al*. (2013) mitochondrial disease (MD) is a group of metabolic disorders with impaired mitochondrial function. Mitochondrial diseases occur when there is a defect in any of the mitochondrial pathways and are classified as a group of disorders that are clinically heterogeneous, and can range from single-organ to multisystem involvement (DiMauro and Schon, 2003; Leonard and Schapira, 2000). The clinical heterogenicity is influenced by a number of factors that include the level of heteroplasmy, type of mutation and affected tissue (Smuts and van der Westhuizen, 2010). Mitochondrial disease is frequent, with an estimated incidence of at least 1 in every 5000 live births and it may include fatty acid oxidation disorders, respiratory chain deficiencies, TCA cycle and pyruvate dehydrogenase complex deficiencies (Schaefer *et al*., 2004; DiMauro and Schon, 2003). Primary mitochondrial disease is the consequence of inherited mutations that result in the dysfunction of the OXPHOS system, while secondary mitochondrial disease occurs when there is a dysfunction of mitochondria due to unrelated genetic or environmental causes (Haas *et al*., 2007).

Mitochondrial disease is associated with a variety of clinical syndromes and in adults they present with the following classical mitochondrial syndromes (Koenig, 2008): mitochondrial encephalomyopathy with lactic acidosis and stroke-like episodes (MELAS), chronic progressive external ophthalmoplegia (CPEO), maternally inherited diabetes and deafness (MIDD), the Kearns-Sayre syndrome (KSS), neurogenic ataxia with retinitis pigmentosa (NARP), Leber hereditary optic neuropathy (LHON), myoclonic epilepsy with ragged-red fibres (MERRF), Pearson syndrome and Leigh syndrome (Pfeffer *et al*., 2012). In children the manifestations are more non-specific (Smuts and van der Westhuizen, 2010)

2.6 Respiratory chain deficiencies

Respiratory chain deficiencies are the largest subgroup of mitochondrial disease (Suomalainen, 2011b). When one of the four complexes that form the RC become impaired, it is generally referred to as a respiratory chain deficiency (RCD). Since the RC is encoded by both nDNA and mtDNA it can theoretically present with any symptom, in any organ and at any age (Rötig and Munnich, 2003). RCDs mostly affect tissue that requires high amounts of energy, and organs most frequently affected are the muscle, brain, heart, retina and cochlea (Liang *et al* 2013). Respiratory chain deficiencies should be considered when there are combinations of unexplained neuromuscular and/or non-neuromuscular symptoms. These symptoms should be progressive and involve unrelated tissues or organs (Koenig, 2008; Munnich *et al*., 2013). Presently there is no curative treatment available for RCDs (Suomalainen, 2011b).

2.7 Diagnosis of mitochondrial disease

Diagnosing patients suffering from mitochondrial disease is a major challenge in mitochondrial medicine as it is clinically heterogeneous with an ever-expanding phenotype (Koene and Smeitink, 2011). Patients with a suspected mitochondrial disease are evaluated using a multi-disciplinary approach including clinical, histochemical, molecular and genetic investigations, as well as biochemical assessment (Rodenburg, 2011). The diagnostic procedure typically involves general non-invasive clinical evaluation, leading to imaging and metabolic screening and lastly to specific genetic and biochemical assays, where biopsies may be required (Haas *et al*., 2007).

Diagnostic criteria have been developed which may be used to guide the diagnosis process, such as the Mitochondrial Disease Criteria developed by Wolf and Smeitink (2002), which is used to guide patient selection, and the Nijmegen Centre for Mitochondrial disorders scoring system. These criteria are based on combining clinical, biochemical and genetic findings (Koenig, 2008).

2.7.1 Clinical basis of diagnosis

Diagnosis begins with a clinical examination, where typical sets of symptoms are searched for (Sumomalainen, 2011). Clinical presentation commonly found in children includes: cardiomyopathy, external ophthalmoplegia, failure to thrive, metabolic encephalopathy, motor regression, ptosis and seizures.

For adults the most common clinical manifestations include: cardiac manifestations, central nervous system involvement, exercise intolerance, endocrine abnormalities, gastrointestinal system abnormalities, muscle weakness, ophthalmological abnormalities and sensorineural hearing loss (Liang *et al* 2013). As mentioned in Section 2.5, mitochondrial disease is associated with classical mitochondrial syndromes that tend to have well established phenotypes, however, in many cases patients present with symptoms that do not meet the criteria for any of these syndromes (Smuts and Van der Westhuizen, 2010).

2.7.2 Minimally-invasive biochemical investigations

Prior to the examination of muscle samples, biofluids (including urine and blood) are usually analysed (Rodenburg, 2011). Generally, metabolic studies include analyses of organic acids, amino acids, pyruvate and acyl-carnitines. These metabolic studies may be very informative, keeping in mind that normal values do not rule out a mitochondrial disease (Koenig, 2008). Proactive clinical testing includes carbohydrate loading followed by the measurement of plasma pyruvate, lactic acid and alanine concentrations. Additionally, fasting studies are often carried out and can reveal a tendency towards secondary defects in fatty acid oxidation or hypoglycemia. The results of these tests can be used as a guide to select the correct diet that could aid in minimizing lactic acid and pyruvate concentrations (Haas *et al*., 2008).

2.7.3 Invasive biochemical investigations

Muscle biopsies are performed to obtain skeletal muscle samples that can be used for biochemical, histochemical, immunohistochemical and morphological studies. Biochemical studies include spectrophotometric assays of enzyme activity and studies that determine the function of intact mitochondria. Enzyme activity measurements can be performed using tissue homogenates or isolated mitochondria. The RC enzyme complex activities can be analysed in isolation as complex I, complex II, complex III or complex IV or alternatively studied together as Complex I+III or Complex II+III. Results of these measurements are reported relative to citrate synthase, a marker enzyme, or as internal ratios where the activity is expressed against Complex II or complex IV (Haas *et al*., 2008). Other methods of examining OXPHOS enzymes include colorimetric gel measurements of enzyme activities after blue-native gel electrophoresis as well as alternative colorimetric assays (Rodenburg, 2011).

Molecular investigations typically screen for known mtDNA point mutations. Various tissues can be used to detect mutations in mtDNA, however large scale rearrangements and deletions are

more likely found in muscle samples (Koenig, 2008). Next-generation sequencing of nDNA is the golden standard for detecting mutations in nuclear genes (Haas *et al*., 2008).

Morphological studies performed on skeletal muscle show a phenomenon called ragged-red fibers (RRF), which is the accumulation of mitochondria in the subsarcolemmal space. These RRF are visualized by Gömoritrichrome staining. Additional histochemical stains that are used for analysing mitochondrial enzyme activity include cytochrome c oxidase, NADH dehydrogenase and succinate dehydrogenase (Haas *et al*., 2008).

2.8 Biomarkers and RCDs

According to Parida *et al*. (2010), a biomarker can be defined as "a characteristic (or metabolite, in the case of metabolomics) that is objectively measured and evaluated as an indicator of a physiological or pathological process or pharmacological response(s) to a therapeutic intervention". The term biosignature refers to a profile of combined biomarkers (Parida & Kaufmann, 2010). Diagnostic tools currently used for RCDs lack sensitivity as well as specificity. Furthermore, skeletal muscle sampling is a very invasive procedure and is prone to complications. The diagnostic process involved in RCDs would benefit from the identification of biomarkers since they can be used as a non-invasive screening tool. For this reason, biomarkers capable of being easily detected in bodily fluids, for instance in urine, are in high demand. Biomarkers for RCDs would be used to support a specific diagnosis and monitor progression and manifestations of the disease, as well as responses to therapy. In order to search for new biomarkers, "omics" approaches such as metabolite profiling are used and by studying transcriptomic, proteomic and metabolomics profiles induced by defects in genes, a global view of the metabolic pathways and biomolecules affected in disease can be seen. Measuring metabolite levels and the variation of biofluids can offer insight into the disease progress (Suomalainen, 2011a, Reinecke *et al*., 2012).

2.9 Metabolomics

According to Dunn *et al*. (2005) metabolomics can be defined as the non-biased identification and quantifications of all metabolites in a biological system by using selective and sensitive techniques. Metabolomics is the final step of the "omics" cascade, which starts with genomics, followed by transcriptomics leading to proteomics and finally metabolomics (Dettmer *et al*., 2007).

Metabolomics is progressively being used in the process of identifying biomarkers that could assist with diagnosis and risk predictions of diseases (Xia *et al.,* 2013). Biomarker discovery with the use of metabolomics is supported by the assumption that firstly, metabolites play a vital role in biological systems and secondly, that a disease causes disruptions in metabolic pathways (Monteiro *et al*., 2013). The metabolic differences between disrupted and undisrupted systems (for instance patients affected by a disease and healthy subjects) found with metabolomics analysis may lead to a better understanding of the underlying pathology (Madsen *et al*., 2010).

The metabolome represents the collection of all metabolites present within a biological system (Dunn *et al*., 2005). The variety of compounds that are found can be divided into different classes such as amino and organic acids, lipids, nucleotides, etc. Metabolites found in biofluids consist of different compound classes that differ in size and polarity, with a wide range of concentrations; thus there is no single analytical approach that has the ability to analyse all the metabolites present in a biological sample (Rhee and Gerszten, 2012). When measuring the metabolome, mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy are the techniques most commonly used (Emwas *et al*., 2013).

2.10 Analytical technologies used in metabolomics

2.10.1 Nuclear magnetic resonance spectroscopy (NMR)

The basic principal of NMR involves atomic nuclei being subjected to a powerful magnetic field and exposed to a radio frequency pulse. These nuclei are capable of spin and align themselves with respect to the magnetic field. The nuclei are then exposed to a radio frequency pulse that excites these atoms to a higher energy state. Upon removal of the radiofrequency, the nuclei relax and return to their original lower energy orientations (Goldsmith *et al*., 2010). NMR has a wide variety of advantages when applied to metabolomics applications such as: no sample pre-treatment required, high throughput capacity as well as excellent repeatability and specificity (Dunn *et al*., 2005). A key benefit of NMR spectroscopy, when doing metabolic profiling studies, is that the technique is quantitative and minimum sample preparation is required, such as separation or derivatization (Smolinska *et al*., 2012). However, NMR is relatively low in analytical sensitivity, meaning that only high-abundance metabolites will be detected while low-abundance metabolites that may be of high value are not detected (Emwas *et al*., 2013).

The spectrum produced by NMR analyses is complex, with signals very close together, possibly resulting in low abundance metabolites being masked by high abundance metabolites (Nikolic *et al*., 2014).

2.10.2 Mass spectrometry

Mass spectrometry (MS) is a widely applied platform used in metabolomics which operates on the basis of forming ions, separating them according to mass-to-charge (m/z) ratio, followed by the detection (Dunn and Ellis, 2005). This platform provides rapid, sensitive and selective quantitative and qualitative analyses (Dunn *et al*., 2005). By using MS, metabolites can be identified using mass spectrum interpretation as well as determination of the molecular formula by using accurate mass measurements. In recent years, there has been large and on-going progress in the field of MS based metabolomics analysis, providing researchers with numerous choices regarding separation techniques and quantification (Lu *et al*., 2008). There are numerous mass spectrometer options including the type of analyser, the configuration, and the combination of different instruments that can be used to form a hybrid mass spectrometer. Types of analysers include quadrupole, time-of-flight (TOF), ion traps and Fourier transform ion cyclotron resonance. Configuration of the mass analysers can be arranged in tandem, such as triple quadrupole mass spectrometer or quadrupole-TOF (QTOF) (Lu *et al*., 2008). MS allows sensitive detection of metabolites and is generally combined with separation techniques such as gas chromatography (GC) and liquid chromatography (LC) or even capillary electrophoresis (Madsen *et al*., 2010). These separation techniques are used to improve the selectivity, sensitivity as well as resolution of MS instruments (Monteiro *et al*., 2013). Direct infusion is an alternative option but it is not preferable due to ion suppression (Dunn *et al*., 2010).

2.10.2.1 Gas chromatography mass spectrometry

Gas chromatography is generally used to separate volatile and thermally stable compounds, after which the eluting compounds are detected (Dunn *et al*., 2011). This is a very popular platform, although it is biased against high molecular weight and non-volatile metabolites. Derivatization procedures allow coverage of a large part of the metabolome. Following derivatization, polar compounds (such as sugars, amines, acids, thermolabile molecules and other high molecular weight compounds) can be detected.

Other advantages of derivatization include improved peak shape and chromatography resolution as well as increased thermal stability of compounds (Monteiro *et al*., 2013). However, derivatization is a time-consuming step that also increases the chance of metabolite loss and results in the formation of multiple products for a single metabolite (Xu *et al*., 2010).

2.10.2.2 Liquid chromatography mass spectrometry

Liquid chromatography (LC) is a widely used platform as it has the ability to simultaneously measure/detect an enormous number of different metabolites (Monteiro *et al*., 2013). Minimal sample preparation is usually needed and derivatization is rarely required. There are a number of ionization options that could be considered when using LC approaches, these include: atmospheric pressure photoionization (APPI), atmospheric pressure chemical ionization (APCI) and the soft ionization technique of electrospray ionization (ESI) (Madsen *et al*., 2010). Separation of metabolites using LC-MS techniques are achieved by the liquid mobile phase and solid or liquid stationary phase. Additionally, separation can be enhanced by utilizing a mobile phase gradient, where the mobile phase consists of a hydrophilic (e.g. water) and hydrophobic (e.g. acetonitrile) solvent. One of the most common LC-MS methods used in global metabolomics approaches is reverse phase LC-MS (RP-LC-MS) that uses C18 columns to separate analytes based on their polarity. When using RP-LC-MS, hydrophilic compounds will elute first, followed by the more hydrophobic compounds, as the organic composition of the mobile phase increases (Dunn *et al*., 2011). Due to the limited capabilities of reverse phase columns to retain highly-polar compounds, researchers make use of alternative columns to overcome this. Normal phase columns are characterized by the use of a polar stationary phase and a mobile phase that is more apolar (Dejaegher and Vander Heyden, 2010). Hydrophilic-interaction chromatography (HILIC) is a combination of normal phase and reverse phase, with an aqueous polar organic solvent mobile phase and polar stationary phase. The elution order of HILIC columns will mimic that of normal phase, with the least polar compounds eluting first, followed by the most polar compounds (Monteiro *et al*., 2013).

2.10.3 Metabolomics approaches

A typical metabolomics experiment involves different steps that include: experimental design, sampling, sample preparation, sample analysis followed by data pre-processing, data processing and finally statistical analysis. A vital part of a metabolomics experiment is experimental design, so that valid statistical conclusions can be produced (Dunn *et al*., 2005).

There are two main approaches in metabolomics, namely, targeted analysis and untargeted profiling (or metabolic profiling). Targeted analysis is hypothesis driven and entails the quantification of a small set of selected metabolites. Untargeted profiling is hypothesis generating and aims to measure as many metabolites in a biological system or sample as possible (Dunn and Ellis, 2005; Dunn *et al*., 2010). The advantage of performing untargeted profiling is that biomarkers, that may not have been predicted or known beforehand, can be found, and a new hypothesis produced. It is also often the case that metabolites not related by pathway or compound class share the same pattern during the disease and are collectively capable of differentiating between health and disease. Both of these approaches can be used in order to search for new biomarkers (Dettmer *et al*., 2007).

2.11 Metabolomics and RCDs

Various studies have applied metabolomics when studying mitochondrial disease using a number of models, including: animal models (Butler *et al.,* 2013; Falk *et al.,* 2008; Leonig *et al*., 2012; Morgan *et al*., 2015; Sin *et al*., 2002; Vergano *et al*., 2014), tissue cultures (Kami *et al*., 2012; Shaham *et al*., 2008; Shaham *et al*., 2010; Sim *et al*., 2002; Vo *et al*., 2007) and human blood and plasma (Falk *et al*., 2008; Esteitie *et al*., 2005; Shaham *et al*., 2008; Shaham *et al*., 2010; Vergano *et al*., 2014). Urine is a popular sample when investigating mitochondrial disease. Studies conducted using urine mainly focused on comparing healthy controls with patients suffering from OXPHOS defects (Barshop, 2004; Esteitie *et al*., 2005). Smuts *et al*. (2013) proposed a biosignature (six amino acids, six organic acids and creatine) for respiratory chain deficiencies after comparing OXPHOS patients to healthy controls.

Venter *et al*. (2015) proposed a biosignature (Table 2.1) that was different from Smuts *et al.,* (2013) in two ways. Firstly, the healthy control group used in Smuts *et al.* (2013) was replaced with a clinical referred control group, since from a diagnostic viewpoint it is important to be able to distinguish a clinical referred control from a RCD patient, as both groups show similar clinical symptoms. Although the clinical referred control group presented with symptoms similar to a RCD, these patients could not be confirmed as an RCD on an enzymatic level. Secondly, an untargeted LC-MS metabolomics approach using both negative and positive electron spray ionization was used. After univariate and multivariate statistical analysis, 12 features were identified that had the ability to distinguish between the clinical referred control and RCD patient groups. This biosignature was a big step in the discovery of a screening tool instrument that could be helpful in selecting patients for muscle biopsies.

Table 2.1: The 12 features of the biosignature for RCDs proposed by Venter *et al*. (2015)

	Annotated name	Mass	Retention time	Chromato-graphy	Ionization
1	AMP	347.0616	2.30	RP	Positive
2	$C_9H_{19}NO$	157.1464	30.3	RP	Positive
3	$C_{23}HNO_8S_4$	546.8560	12.7	RP	Positive
4	Unknown 1	1043.4677	11.97	RP	Positive
5	N-succinyl-L-L2.6 diaminopimelate	290.1126	26.9	RP	Positive
6	$C_{16}H_{26}O_4S$	314.1547	24.9	RP	Positive
7	Fragment of tetradecanedioic acid	136.1248	27.9	RP	Positive
8	$C_{14}H_{22}O_2$	222.1615	24.7	RP	Positive
9	$C_{14}H_{24}O_2S$	256.1491	27.67	RP	Positive
10	$C_{20}H_{29}NO_{10}$	443.1786	21.7	RP	Positive
11	Oxoglutaric acid	146.0215	2.35	HILIC	Negative
12	N-Acetyl asparagine	174.0641	8.7	HILIC	Negative

2.12 Problem statement, aim and objectives

Diagnosing mitochondrial disease is very difficult due to the diverse symptoms and onset periods of the disease. Enzyme analyses on muscle samples, obtained from a muscle biopsy, remain the gold standard when it comes to confirming RCDs. However, obtaining a muscle biopsy is a daunting task that involves hospitalization, anaesthetics and trauma. Another challenge that clinicians face, is the selection of patients who must undergo a muscle biopsy. For this reason, other less invasive options ought to be explored to aid in patient selection. Biomarkers that could easily be detected in biofluids, such as blood and urine, would be ideal to screen patients with a possible RCD.

At the North-West University (Potchefstroom Campus), a number of studies have been conducted using metabolomics to search for urinary RCD biomarkers. Recently, a biosignature was constructed (consisting of 12 features) where only the accurate mass and retention time of the features was reported (Venter *et al.,* 2015). Since the true identity of the 12 features remained unknown, this biosignature still needed to be verified using the limited information available. Additionally, this biosignature had the potential to be improved since only a single platform, LC-MS, was used to construct it. By using GC-MS and NMR, a larger portion of the metabolome would be covered, potentially yielding additional discriminative features.

Expanding the existing biosignature, in terms of metabolome coverage, could also reveal metabolite covariance structure with a greater discriminative power than the current biosignature.

Therefore, the **aim** of this study was to improve the RCD biosignature proposed by Venter *et al*. (2015).

The **objectives** formulated to achieve the aim of this study were divided into two separate phases, namely, a verification phase and expanding phase. The specific objectives were:

Verification phase

i. Verify the true existence of the 12 features of the proposed biosignature by ascertaining whether that they are not artifacts from the analysis or data processing procedures.
ii. Characterize the 12 features by identifying the small chemical entities or substructures responsible for the characteristic masses and retention times.

Expanding phase

iii. Analysis of an identical sample cohort with a selected GC-MS method and identification of discriminative features.
iv. Analysis of an identical sample cohort with NMR and identification of discriminative features.
v. Compilation of an improved biosignature, consisting of the verified LC-MS features, as well as new discriminating features found using the GC-MS and NMR analytical platforms.

2.13 Experimental approach

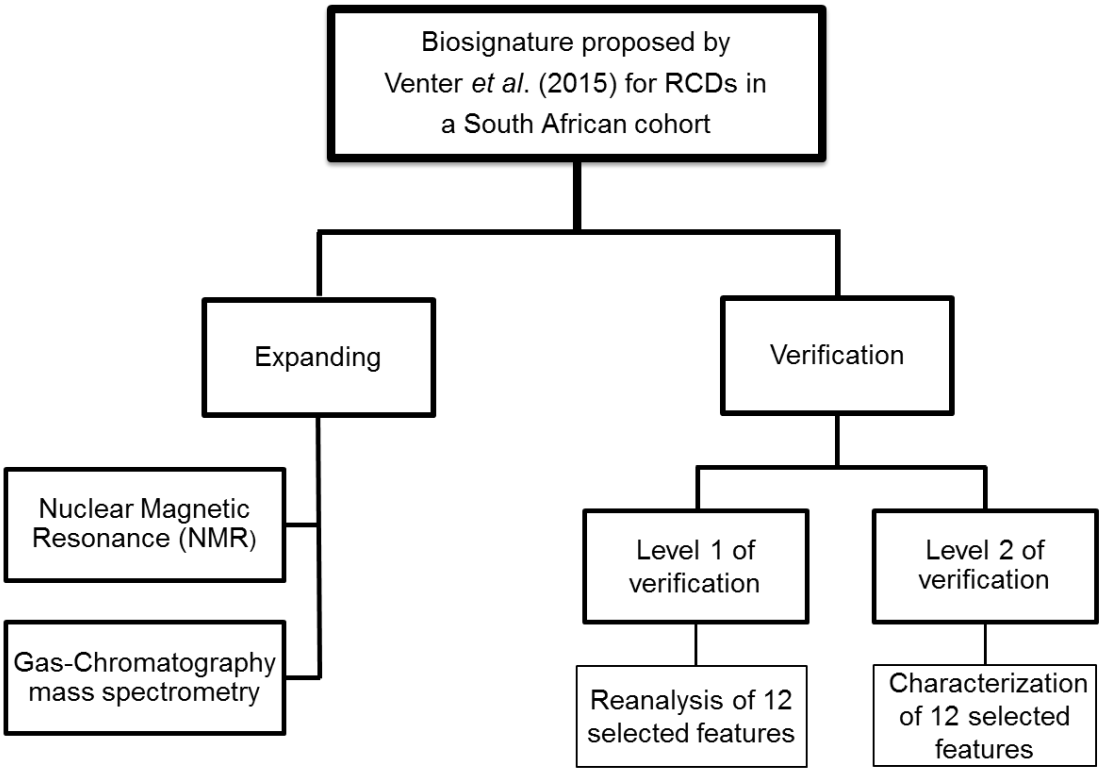

Figure 2.5: Experimental approach workflow. The experimental approach of this study consisted of two phases, a verification phase and an expanding phase. As part of the expanding phase, NMR as well as GC-MS analysis was used. The verification phase was more complicated and consisted of two levels, namely, reanalysis and characterization of the 12 selected features.

As shown in Figure 2.5, the experimental approach of this study consisted of two phases, a verification and expanding phase. The first level of verification involved the analysis of an identical sample group, used in Venter *et al*. (2015), on an alternative analytical platform (LC-QTOF) using a targeted MS/MS approach. This involved analysing a pooled urine sample, searching for the selected features using accurate masses and retention times (as reported by Venter *et al*. (2015)) followed by analysis of the raw data using alternative software. The second level of verification, which is also linked to the first level, was to characterize the features, i.e. to identify the small chemical entities that gave the characteristic masses and retention times. Characterization of the features was performed by analysing the pooled urine sample and using additional analytical techniques in order to generate spectra that could lead to a possible identification.

The expanding part of the study consisted of GC-MS and NMR analyses on the identical sample group (Venter *et al.,* 2015), including method evaluation and sample analysis of both platforms, followed by data analysis.

CHAPTER 3

VERIFYING AND CHARACTERIZING THE PROPOSED LC-MS BASED RCD BIOSIGNATURE

3

3.1 Introduction

The urinary biosignature for RCDs proposed by Venter *et al*. (2015) consists of 12 features. Six of these features are annotated to a metabolite name, four features are reported with a molecular formula, and for the two remaining features, only accurate mass and retention times are reported. Thus, the true identity of the 12 features remains elusive. The term *feature* is used in this section to describe a two dimensional (retention time, RT, and mass-to-charge ratio, m/z) LC-MS signal (Tautenhahn *et al*., 2008). Verification of this biosignature was thus necessary and required reanalysis of the same sample cohort on the identical (or different) instrument(s) and analysis of the data with the identical, or a different, software package. Since the true identities of the features had not been confirmed, the features could be a metabolite, a fragment of a metabolite, an artefact of analysis (such as false peaks related to detector ringing) or an artefact of data processing (due to over-deconvolution). The term *metabolite* that is used in this study refers to all small molecule compounds that are measured in urine and can be derived from any source (including diet, drugs or environmental exposure).

In order to verify the proposed biosignature, all 12 features were evaluated using the first two levels of verification. The first level of verification was performed by reanalysing the 12 selected features using a more targeted approach on an alternative analytical instrument (LC-QTOF). Additionally, the data was reanalysed with a different software program. If the features could be detected using this approach, it would confirm that the features were indeed chemical entities and not artefacts of analysis and data processing. Thereafter, the features of interest were reanalysed in an attempt to characterize the chemical entities responsible for the characteristic masses and retention times; this was part of the second level of verification. By connecting the features to a certain metabolite class, one is also verifying that the feature is true. The term *true feature* refers to a detectable compound belonging to a metabolite class. However, the possibility exists that some of the features found in the proposed biosignature (Venter *et al*., 2015) were novel metabolites or metabolites not available in the databases that were used to annotate features. Therefore, the absence of an identity does not mean that the feature does not exist.

This chapter includes all the techniques used and data obtained, in an attempt to verify and characterize the features of the biosignature proposed by Venter *et al.* (2015).

3.2 Materials and methods

3.2.1 Reagents

Spectrometry-grade acetonitrile (Honeywell Burdick & Jackson, Cat # BJ015CS) and water (Cat # BJ365CS) were purchased from Anatech. Formic acid (Cat # 06440) and acetic acid (cat # 49199) was purchased from Sigma-Aldrich.

3.2.2 Internal standards

The internal standards nor-leucine (Cat # 74560), 2-acetamidophenol (Cat # A700025GA), caffeine (Cat # 44818) and 3-phenylbutyric acid (Cat # 116807) were purchased from Sigma-Aldrich. Internal standard stock solutions with concentrations of 1000 ppm were prepared in Millipore (MilliQ) water. An internal standard mixture was prepared by combining 200 µL of each standard stock solution. This resulted in a final internal standard mixture with a concentration of 333.3 ppm for each of the three internal standards. To each sample, 30 µL of this internal standard mixture was added, dried and later dissolved in 100 µL mobile phase, resulting in a final concentration of 100 ppm for each internal standard (nor-leucine, 2-acetamidophenol and caffeine, respectively).

3.2.3 Biological samples

3.2.3.1 Ethical approval

Ethical approval for this study was obtained from the Ethics Committees of the University of Pretoria (No. 91/98 and amendments) and the North-West University (NWU-00170-13-S1). The parents of all the patients, as well as controls, gave informed consent for the use of their children's muscle biopsies and urine samples for research purposes.

3.2.3.2 Patient selection

Sample collection took place at the Steve Biko Academic hospital's Paediatric Neurology Unit in Pretoria, South-Africa. Muscle biopsies from the *Vastus lateralis* were performed on patients who had a Mitochondrial Disease Criteria score ≥ 6 (Wolf and Smeitink, 2002), or who presented with a classical phenotype that was related to a mitochondrial disorder and involved two or more distinct physiological systems. Enzyme activities of RC enzymes (Complex I – Complex IV), pyruvate dehydrogenase complex (PDHc) and citrate synthase were measured in muscle, as described by Smuts *et al*. (2010). The results of enzyme analyses were used to classify patients in either the RCD or CRC groups. The criteria for a RCD is defined as when an enzyme activity is lower than the reference values expressed against at least two of the three enzyme markers (CS, Complex II and Complex IV). Any deficient marker should be excluded from interpretation, and if both Complex II and Complex IV are deficient, only CS should be used for identification (Smuts *et al*., 2010). Urine samples of patients were collected and stored at -80°C for the metabolomics investigation.

3.2.3.3 Experimental groups

The experimental samples used in this study consisted of two separate groups, a RCD patient group and CRC group (Figure 3.1). The RCD group consisted of patients with a confirmed RCD on the enzyme level. The second group was a control group that presented with clinical symptoms similar to RCDs and had undergone muscle biopsies, but where no RCD could be detected on an enzyme level.

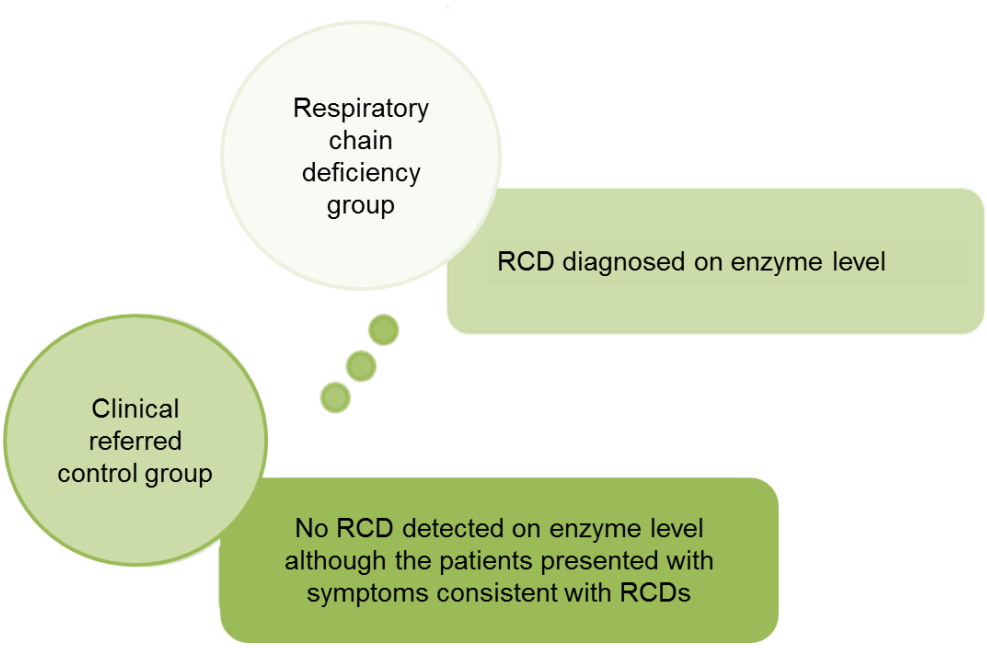

Figure 3.1: Experimental groups used in the study. The two experimental groups used in this study, which included a respiratory chain deficient (RCD) group and clinically referred control (CRC) group.

In the study by Venter *et al.* (2015), a total of 46 urine samples were analysed to construct the urinary biosignature. In this study, the same urine samples were used to validate the proposed biosignature. However, due to the limited sample volume available, not all the samples could be included in this study. Therefore, 12 samples (ten RCD patients and two CRC samples) were excluded due to sample availability, yielding a total of 34 samples used in this study (Table 3.1).

Table 3.1: Classification of experimental samples

Experimental group	Gender		Race	
	Male	Female	African	Caucasian
RCD patients	8	13	15	6
CRC	5	8	8	5

Table 3.1 illustrates the classification of patients that formed part of this study. The RCD group consisted of a total of 21 patients, of which eight were male and 13 female, when classified by gender. By race, 15 patients were African and only six were Caucasian. The CRC group included a total of 13 patients, five male and eight female, when classified by gender. Classified by race, the group consisted of eight African and five Caucasian.

3.2.4 Preparation of the pooled urine sample

A pooled urine sample was prepared by mixing an equal volume of all the RCD patient urine samples, and it was then aliquoted and frozen at -80°C. This pooled urine sample was used in the study for the verification of the biosignature proposed by Venter *et al*. (2015). The creatinine value of the pooled urine sample was determined by Ampath Laboratories, Potchefstroom, South Africa. A predetermined volume of urine (containing 0.25 µmoles creatinine) was placed in a microcentrifuge tube along with 30 µL of the internal standard solution and 20 µL acetonitrile. The sample was frozen overnight at -80°C and then freeze dried using a Savant SpeedVac SVC 100 concentrator. The dried sample was then re-dissolved in 100 µL water by incubating it for 30 minutes at room temperature followed by vortexing for 10 seconds. The next step was to centrifuge the sample at 25 055 X *g* for 10 minutes at 4°C, followed by transfer of the supernatant to a vial fitted with a pointed insert.

3.2.5 Chromatography

3.2.5.1 Reverse phase

The same reverse phase chromatographic conditions were used as described by Venter *et al.* (2015). For separation, an Agilent Zorbax SB-Aq reverse phase column (3.5 µm, 2.1 x 150 mm) was used. The mobile phases of choice (Section 3.2.1) were 0.1% formic acid in water for the aqueous phase (solvent A) and 0.1% formic acid in acetonitrile for the organic phase (solvent B). The gradient used during the chromatographic run started at 100% solvent A for the first 5 minutes, followed by an increase to 35% of solvent B over a period of 25 minutes. Thereafter, the gradient was increased to 70% solvent B at 35 minutes and finally 100% at 36 minutes. The gradient was maintained at 100% solvent B for 3 minutes followed by a decrease to 0% within 3 minutes. To ensure equilibration of the column, a post run was introduced that lasted 8 minutes, resulting in a total run time of 50 minutes (although the mass spectrometer only acquired data for 42 minutes). With a constant column temperature of 30°C, 5 µL of the sample was injected and a constant flow rate of 0.2 mL/min was maintained throughout the run.

3.2.5.2 Hydrophilic interaction chromatography

The same normal-phase chromatographic conditions were used as described by Venter *et al*. (2015). An Agilent ZORBAX RRHD HILIC plus (Chemetrix, 959759-901) column (1.8 µm, 2.1 X 150 mm) fitted with a Phenomenex guard column (AFO-8497) was used. The mobile phases of choice (Section 3.2.1) were 5 mM acetic acid in water (solvent A) and 5 mM acetic acid in acetonitrile (solvent B). The total chromatographic run was 25 minutes with a post run of 10 minutes. The gradient used during the chromatographic run began at 10% solvent A which was maintained for 2.5 minutes, after which the gradient was linearly increased to 50% solvent A at 10 minutes. For a total of 5 minutes, solvent A was kept at 50% before the gradient was lowered back to 10% solvent A at 20 minutes. Finally, the gradient was kept at 10% solvent A for 5 minutes until the run was complete. A post run of 10 minutes was used to ensure column regeneration. The column temperature was set at 35°C and a flow rate of 0.25 mL/min was maintained. A total of 5 µL sample was injected.

3.2.6 Instrumentation

Different instruments were used to verify and characterize the biosignature. The first was an Agilent 1200 series LC system coupled to a 6510 Q-TOF mass analyser (Agilent Technologies, Santa Clara, CA, USA) consisting of a Micro Vacuum Degasser; Binary pump SL; Preparative Autosampler HiP-ALS SL; Thermostat ALS and Thermostatted Column Compartment SL. The drying gas temperature was set at 280°C, with a flow of 8 L/min and nebulizer pressure of 30 psi. The dual ESI source was set up for positive ionization mode and negative ionization mode, respectively. Mass drift was monitored and corrected for by constantly injecting a reference mix with masses 121.050873 (protonated purine) and 922.009798 (protonated hexakis (1H, 1H, 3H-tetrafluoropropoxy).phosphazine or HP-921) m/z for positive ionization and 119.036320 (proton abstracted purine) and 966.000725 (formate adduct of HP-921) m/z for negative ionization.

Fractionation of samples were performed on an Agilent 1200 liquid chromatography system (Agilent Technologies, USA), equipped with a quaternary pump, an online degasser, and a column temperature controller, fraction collector, coupled with an diode array detector. The analytical column was a ZORBAX SB-Aq reverse phase C18 column (3.5 µm, 2.1 X 150 mm) and the column temperature was kept at 30°C. The gradient used for chromatographic separation is described in Section 3.2.5.1. The flow rate was 0.2 mL/min, and the injection volume was 5 µL.

Samples were analysed using an Agilent 1200 LC coupled to an LC/MSD Trap XCT electrospray ion trap mass spectrometer. Source settings used for the ionization were as follows: nebulizer gas flow, 30 psi; dry gas flow, 9.00 L/min; electrospray voltage of the ion source, 3000 V; capillary temperature, 280°C; capillary exit, - 158.5 V; skimmer, 40 V. Nitrogen (>99.99%) and Helium (>99.99%) were utilized as drying and collision gas, respectively. The full scan of ions ranging from m/z 100 to 1000, in the positive mode, was carried out.

Lastly the Synapt G2 Si (hybrid IM-MS system) (Waters) instrument was used with an Acquity UPLC system equipped with ESI source operating in negative mode. The nebulization gas was set to 600 L/min at a temperature of 500°C, and the cone gas was set to 20 L/min. The capillary voltage and cone voltage were set at 2760 V and 30 V, respectively. The Q-TOF acquisition rate was 0.1 s, with argon used as the collision gas at a pressure of 7.066×10^{-3} Pa. The energies for collision-induced dissociation (CID) were set at 10 and 40 eV for the fragmentation information in low (first channel) and high (second channel) energy respectively. All MS data was acquired using the lock spray to ensure accurate and reproducible mass. Leucine-enkephaline was used as the lock mass (m/z 556.2771) in positive ionization mode.

3.2.7 Fraction collection and preparation of fractionated sample

Since urine samples contain a vast number of metabolites and are therefore very complex, the decision was made to fractionate the pooled urine sample in order to obtain numerous fractions. The idea was to use a specific fraction that corresponded to the retention time of a given feature, to search for said feature and to attempt to characterize the feature. In theory, the fractions obtained should be less complex (contain less metabolites) and allow for easier feature characterization than the original urine sample (which contained hundreds to thousands of metabolites). Also, by repeating the fraction collection multiple times, one can obtain numerous fractions of the same retention time. By combining these fractions, drying them and then dissolving the combined fraction in a small volume of mobile phase, the fraction is actually concentrated and the concentration of the feature of interest is increased, possibly making it easier to detect such a feature while attempting to characterize it.

Time-based fraction collection was performed with fractions collected at every minute over the 42 minute reverse phase chromatographic separation (Section 3.2.5.1), thus dividing the pooled sample (Section 3.2.4) into 42 fractions. A diode array detector was used to monitor fraction collection by detecting caffeine in fraction 24 at a wavelength of 273 nm (to ensure that retention time did not drift). The fractionation experiment was repeated 35 times, thus 35 volumes of the 42 fractions were collected. The 35 volumes of a given fraction were combined,

dried and dissolved in 100 µL water:acetonitrile (50:50). Thus the factor by which the fraction was concentrated, was seven times that of the original urine sample.

3.2.8 Relative retention time calculation

Since a different LC-QTOF (Agilent 1200 series LC system coupled to a 6510 Q-TOF mass analyser) was used in this study, compared to the one used by Venter *et al*. (2015) (Agilent 1290 series LC system coupled to a 6540 Q-TOF mass analyser), retention time drifts occurred. This was to be expected due to the different types of tubing in the two LC systems. To overcome this, relative retention times (also referred to as a *retention time index*) had to be determined for all 12 features by using the retention time of the internal standard that eluted closest to the specific feature. To achieve this, an internal standard sample, containing nor-leucine, 2-acetamidophenol and caffeine (3-phenylbutyric acid in the case of negative ionization), were firstly analysed. Figure 3.2 shows a representative chromatogram of the internal standard mixtures analysed in positive ionization mode. Venter *et al*. (2015) reported the following retention times for each internal standard: nor-leucine at 2.4 minutes, 2-acetamidophenol at 12.3 minutes and caffeine at 18.4 minutes.

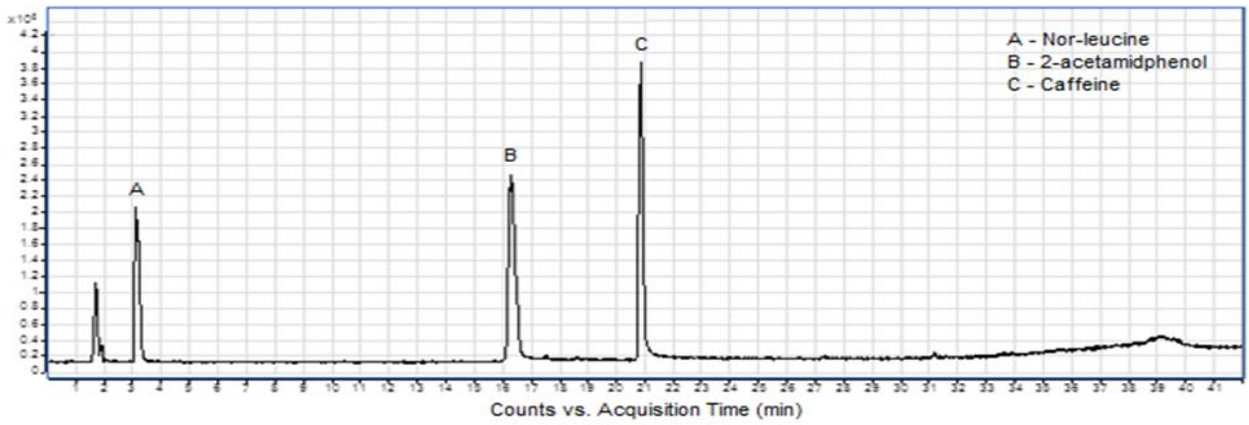

Figure 3.2: The total ion chromatogram obtained for the three internal standards in positive ionization mode. As indicated in the figure, nor-leucine (A) elutes first (3.1 minutes), followed by 2-acetamidophenol (B) (16.3 minutes) and then caffeine (C) (20.4 minutes).

Using this information, the relative retention times were calculated using Equation 3.1 and the results are shown in Table 3.2.

Equation 3.1: Relative retention time calculation

$$\text{Relative retention time} = \frac{\text{RT (old) x RT internal standard (new)}}{\text{RT internal standard (old)}}$$

RT (old) = retention time of the features as reported by Venter *et al.* (2015).
RT internal standard (old) = retention time of the internal standard as reported by Venter *et al*. (2015).
RT (new) = retention time of the internal standard as measured in this study.

It is important to note that the fraction collector also uses a different LC system, and thus also results in retention time drifts. For this reason, the fractions do not correlate with the relative retention times due to the different LC system of the fraction collector and LC-QTOF. In Table 3.2, the relative retention time and fraction (containing each feature) is listed along with feature number, annotated name and mass. The features of the biosignature are listed as Features 1-12 where the annotated names provided by Venter *et al*. (2015), as listed in Table 2.1, were replaced with a feature number, in order to avoid confusion.

Table 3.2: Features of the biosignature proposed by Venter *et al*. (2015).

Feature	Annotated name (Venter *et al*., 2015)	Mass-to-charge (m/z)	Retention time (Venter *et al*., 2015)	Relative retention time	Fraction
Feature 1	AMP	348.0686	2.3	2.97	3
Feature 2	$C_9H_{19}NO$	158.1537	30.3	33.76	38
Feature 3	$C_{23}HNO_8S_4$	547.8636	12.7	16.52	19
Feature 4	Unknown 1	522.7411	11.97	15.57	18
Feature 5	N-succinyl-L-L2.6 diaminopimelate	291.1197	26.9	29.97	34
Feature 6	$C_{16}H_{26}O_4S$	332.1886	24.9	27.742	32
Feature 7	Fragment of tetradecanedioic acid	137.1321	27.9	31.08	36
Feature 8	$C_{14}H_{22}O_2$	223.1688	24.7	27.52	30
Feature 9	$C_{14}H_{24}O_2S$	257.1563	27.67	30.83	35
Feature 10	$C_{20}H_{29}NO_{10}$	444.186	21.7	24.18	28
Feature 11	Oxoglutaric acid	145.0215	2.35	-	-
Feature 12	N-Acetyl asparagine	173.0641	8.7	-	-

(-) Analysis was not performed

3.3 Verification of the biosignature features

The biosignature was verified on the first two levels of verification. Each level involved various techniques and methodologies which will be discussed in detail in the following sections.

3.3.1 First level of feature verification

The first level of verification involved the reanalysis of the existing data on a different software package, followed by analysis of a sample that represented the original sample cohort (pooled urine sample in this case), on an alternative analytical instrument. By doing this, the true existence of the features can be verified (Schultz *et al*., 2013). Figure 3.3 provides a summary of the instruments and software that were used as part of the first level of verification.

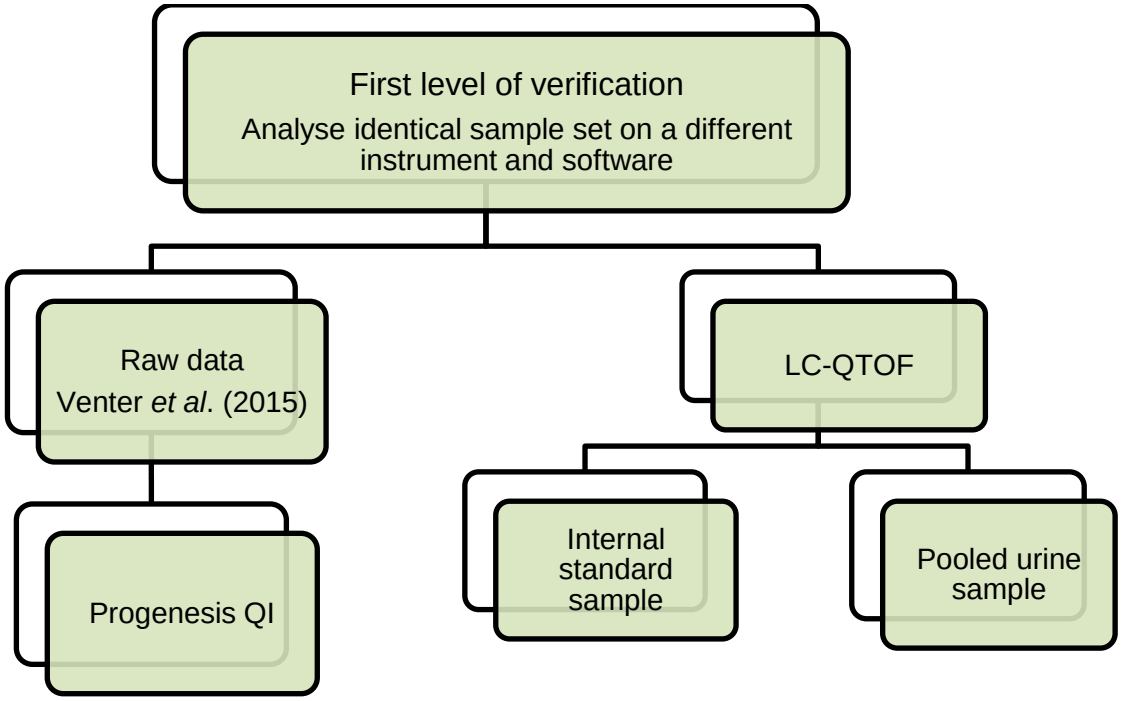

Figure 3.3: The first level of feature verification. The first level of verification entailed the reanalysis of the 12 features using the identical sample set and alternative instrumentation and software. The raw data from the Venter *et al*. (2015) study was reanalysed using Progenesis QI. An alternative LC-QTOF was also used to analyse an internal standard sample (mixture of internal standards), and a pooled urine sample, with more targeted settings.

3.3.1.1 Data analysis with Progenesis QI

It is commonly accepted that many p peaks/features are included in untargeted metabolomics data sets, and while the statistical analysis and data pre-processing (cleaning) should exclude false features, this is not always the case (Chong *et al*., 2015).

Venter *et al*. (2015) used Agilent's Mass Hunter Qual and Mass Profiler Professional software for data processing. In this study, Progenesis QI (Nonlinear Dynamics) was used and was

beneficial in two ways; firstly, as an alternative data processing program it could verify the existence of the features by using different algorithms. If the features (with identical mass) were found at identical retention times, it would confirm that the features were in fact true and that they were not artefacts due to data processing. Secondly, Progenesis QI software is less automated than Mass Hunter, since parameters for alignment and deconvolution can be specifically selected which can lead to errors.

The raw Agilent data files were imported into the Progenesis QI software program where peak detection, deconvolution and alignment of the data was performed. Default peak picking and deconvolution settings were used. A QC sample (data file from Venter *et al.,* 2015) was used as a reference for alignment of the peaks in the other samples, when data extraction was performed in Progenesis QI. Biological samples are complex and contain co-eluting peaks. *Deconvolution* is the process used to ensure separate detection and quantification of these peaks (Fiehn *et al*., 2000). In other words, deconvolution separates peaks that co-elute and then adds (deconvolutes) chemically related ion signals belonging to the same compound. Deconvolution errors can occur due to software, thus leading to artefacts that are not true features (Styczynski *et al*., 2007). A useful characteristic of the Progenesis QI software is that deconvolution can be reviewed and changed if an error is suspected.

In the list of compounds obtained from Progenesis QI, the retention times were used to guide the process of detecting the features together with their accurate masses. Once the features were found, the deconvolution was reviewed, the software provided information regarding the total number of adducts detected for the specific compound, as well as information regarding adducts that could be formed. One metabolite can form multiple ion types, for example $[M+NH_4]^+$, $[M+Na]^+$, $[M+K]^+$ (called adduct ions), in addition to the standard protonated form of the metabolite $[M+H]^+$, when ESI is used (Dunn *et al*., 2013).

The aim of using different software packages was to verify whether or not the feature could be detected when processing with a different program that uses different algorithms to process the raw data. In Progenesis, the relative intensities of adducts are plotted on the neutral mass scale and chromatograms for adducts that indicate the relative intensities are plotted against retention time. Using these figures, the additional adducts not automatically deconvoluted into the compound, can be reviewed. Adducts that are detected, are deconvoluted into one compound when both the mass spectra and chromatograms align.

3.3.1.2 Targeted analysis with LC-QTOF

Another aspect of the first level of verification was to find the features of the biosignature in a pooled urine sample (Section 3.2.4) using a different instrument with more targeted settings, to verify that the feature was actually true and not an analytical artefact. Tandem MS (MS/MS) was used, which only allows the ions of a given m/z (mass-to-charge ratio) to be transmitted past the quadrupole, while the remaining ions are filtered out. The selected precursor ions undergo collision induced dissociation (CID) and the resulting product ions are detected (Chernushevich *et al*., 2001). In order to search for features in the pooled urine sample using their accurate masses, only single mass spectra were required, which meant that the collision energy was kept low to avoid fragmentation. Additionally, time windows were used to search for features in certain time frames (dictated by retention time reported by Venter *et al.* (2015).

An internal standard mixture was prepared to determine the retention times of the internal standards and to calculate relative retention times for features (Section 3.2.8). The internal standard mixture (30 µL) was placed in a microcentrifuge tube together with 70 µL of HPLC-grade water, mixed and transferred to a vial with an insert. The final concentration of each standard was 100 ppm (Section 3.2.2). Since the internal standard's retention times and fragmentation spectra are available, an internal standard sample was also used to monitor targeted MS/MS settings. Thereafter, a pooled urine sample (containing the same internal standards) that represented the sample set was analysed in targeted MS/MS mode with reverse phase chromatography (Section 3.2.5.1) using the LC-QTOF. Urine samples were prepared as described in Section 3.2.4. For negative ionization, the retention times of internal standards were similar to the times Venter *et al*. (2015) obtained, with some minor deviations. Only two features of the proposed biosignature were discovered using negative ionization. For this reason, relative retention time calculation was unnecessary, as larger windows were used to compensate for any time drifts.

3.3.2 Second level of verification

The second level of verification (also linked to level 1) involved characterization of the features, i.e. identification of the small chemical entities that gave the characteristic masses and retention times as seen in Table 3.2. Characterization of the features was performed by using additional analytical techniques. By connecting a feature to a specific metabolite class, one is also verifying that the feature is true and not an artefact of analysis. However, it should be mentioned that novel metabolites can especially be produced in certain diseases and therefore the absence of identity does not necessarily indicate that the feature is false. Figure 3.4 provides a summary of the instruments that were used as part of the second level of verification.

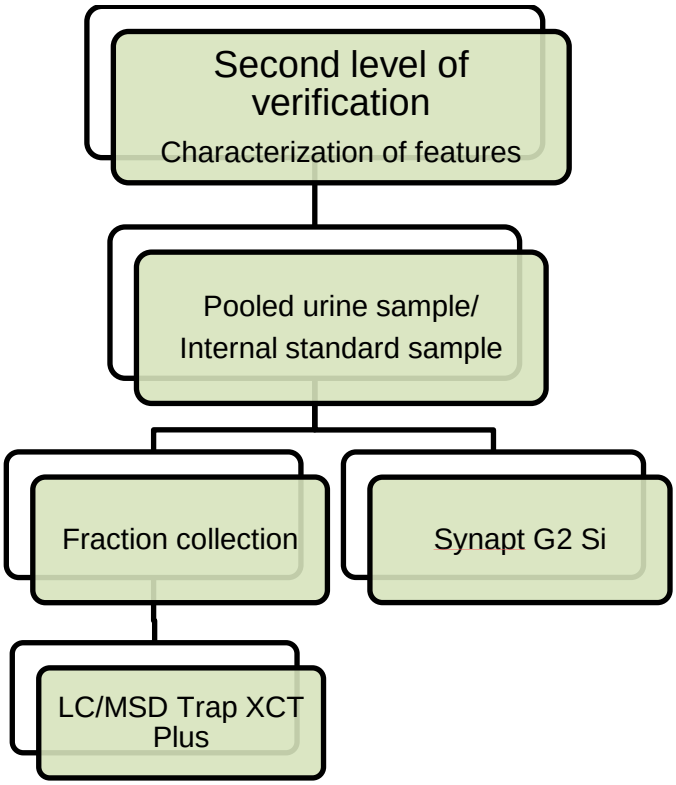

Figure 3.4 : **The second level of verification of the 12 features**. The second level entailed the procedures followed in order to characterize the 12 features of the proposed biosignature. An internal standard mixture (Section 3.2.3) and a pooled urine sample (Section 3.2.4) were analysed using the Ion trap MS (Section 3.2.8) and Synapt MS (Section 3.2.8), with fraction collection (Section 3.2.7) as a concentration step prior to analysis on the Ion trap MS.

The online human metabolome database (HMDB) and METLIN databases were used to identify the features of the proposed biosignature using the accurate mass, isotope ratios and salt adduct patterns, in the study conducted by Venter *et al*. (2015) (Table 3.2).

Features 1, 5, 7, 11 and 12 were annotated to a metabolite name by Venter *et al*. (2015). Characterization of these features would confirm the potential identification already reported by Venter *et al*. (2015), or yield a new identification. A useful tool of Mass Hunter software is its ability to generate/calculate a molecular formula using accurate mass measurements (Kind and Fiehn, 2006). This was done for Features 2, 3, 6, 8, 9 and 10. One remaining feature (Feature 4) could not be annotated to a metabolite or generate a molecular formula and was listed an Unknown 1. Characterization could provide more information on the remaining features.

Metabolomics Standards Initiative (MSI) proposed the minimum reporting standards for chemical analysis in metabolomics (Sumner *et al*. 2007). These levels are not to be confused with the levels of verification used in this study. These levels are linked to level 2 of verification where characterization of metabolites was performed. Four different levels of metabolite identification were described. These levels included: Level 1 - Identified metabolites; Level 2 - Putatively annotated compounds; Level 3 - Putatively characterized compound classes; and Level 4 - Unknown compounds. These levels have been refined since findings often fitted 'in between' these levels. Schymanski *et al*. (2014) proposed the following levels: Level 1 - Confirmed structure; Level 2 - Probable structure; Level 3 - Tentative candidates; Level 4 - Unequivocal molecular formula; and Level 5 - Exact mass. Figure 3.5 provides a summary of the identification confidence levels.

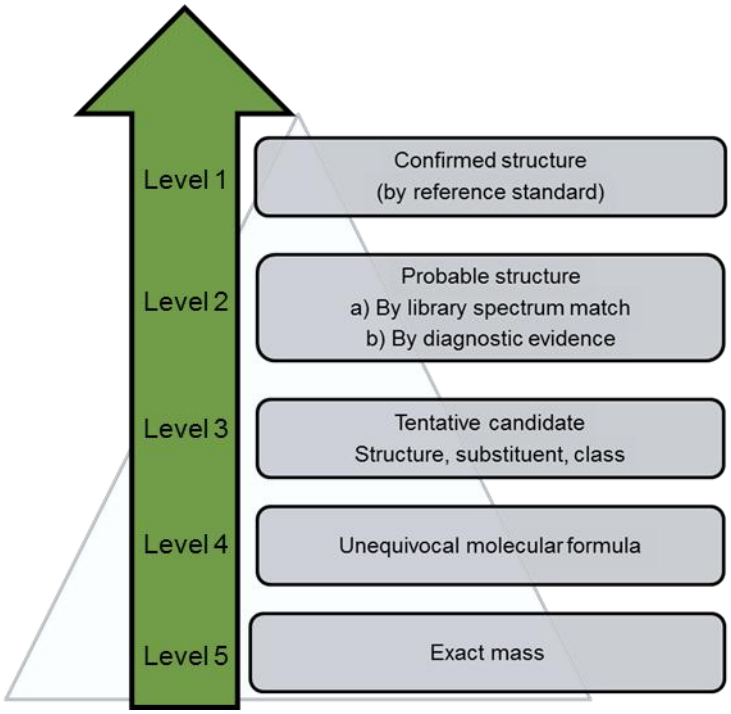

Figure 3.5: Levels of identification confidence. Identification confidence levels proposed by Schymanski *et al*. (2014). (Adapted from Schymanski *et al*., 2014)

The state of the biosignature reported by Venter *et al*. (2015) was on level 4 and 5 according to the criteria of Schymanski *et al.* (2014). The accurate mass and retention time of each feature was reported, as well as the possible molecular formula in some cases. In order to reach at least Level 3, MS^2 was required in an effort to elucidate the structure of the features. The ultimate goal would be to identify the features of the biosignature (level 1), however, identification of metabolites is challenging and time consuming due to the complexity of the metabolome and wide variety of instrumental techniques, as well as software applications, needed (Alonso *et al*., 2015; Clarke *et al*., 2011). This being said, the aim of the second level of verification of this study was to characterize the features of the biosignature, in an attempt to gather more information and to confirm the feature's existence.

It should be noted that for this part of the study no negative ionization analyses were performed since during the first level of verification, only one feature was detected using the LC-QTOF. Due to time constraints, it was decided to mainly focus on the positive ionization that contained most of the features.

3.3.2.1 LC/MSD Trap XCT Plus analyser

The principal of an ion trap mass spectrometer is that it allows the isolation of a single ion in the trap, fragmentation of that ion and then scanning of the product ions. An advantage of the ion trap is that it can form multiple stages of MS (MS^n), which could be useful in identification of features (Prakash *et al*., 2006)

The fractions collected from the pooled urine sample (Section 3.2.6) were used in an effort to characterize the features. Each fraction was directly infused into the ion trap. Firstly, the internal standards were searched for to determine the fraction in which they were contained, and from this it was then possible to predict which fraction a specific feature was likely to be in. Fragmentation spectra for the three internal standards were collected to evaluate the approach, after which nominal mass fragmentation spectra were generated for all the features that could be found. Each particular feature was searched for in multiple fractions to avoid missing a possible detection. The ion trap data was compared to the fragmentation spectra in the METLIN database.

3.3.2.2 Synapt G2 Si hybrid IM-MS system

An advantage of using the Synapt G2 (hybrid IM-MS system) over the other LC-QTOF instruments, is that it is able to provide both low and high energy (fragmentation) data simultaneously, in accurate mass format, and give curated spectra of all the features based on ion mobility (Shah *et al*., 2013). For example, if a compound is fragmented with the Q-TOF, the spectra obtained will often contain larger masses which could be from contaminant peaks, or rearrangements of fragments in the mass spectrometer. The software groups the masses together based on time, and if they were detected together, then it is assumed that they belong together and originated from the same molecular entity. This is often not the case as contaminant and co-eluting peaks are almost always present in biological samples. This problem can potentially be overcome with the Synapt MS, as fragments are grouped based on time and collision cross section (CCS) values (ion mobility data). CCS is able to exploit characteristics of ions in gas phase that are related to their chemical structure and three dimensional orientations. The mobility of ions are affected by CCS and as they move through the neutral gas (under the influence of a specific electrical field), ions are then separated using ion mobility spectrometry. Ions with known CCS values are used to calibrate the ion mobility spectrometer, and from this, the drift times that are measured for unknown ions can be converted to CCS values (WATERS, 2013).

3.3.2.2.1 Sample analysis

The pooled urine sample (described in Section 3.2.3) was analysed in an attempt to produce spectra that could aid in the characterization of the features. Samples were analysed in HDMS^e mode which includes ion-mobility with low and high energy data collection. The data extraction programs used to generate fragmentation spectra were MassLynx (Waters) and MSE viewer (Waters) along with Progenesis QI, and various databases were used for identification purposes.

3.4 Results and discussion

To avoid repetition in this dissertation, the detailed discussion regarding the techniques and methodology that were used to verify the features is confined to a single feature (Feature 9), followed by a discussion of the results obtained for the remaining features.

3.4.1 Verification of Feature 9

3.4.1.1 Data analysis with Progenesis QI

As seen in Table 3.3, Feature 9 was found in the compound list (Progenesis QI) at retention time 27.6 which correlates with the retention time reported by Venter *et al.* (2015) (27.67 minutes). Deconvolution was reviewed, and a precursor ion and adduct ion were detected as 257.1563 (M+H) and 274.1827 ($M+NH_4$) respectively.

Table 3.3: Feature 9 found in raw data when processed by Progenesis

Feature	Mass	Retention Time (minutes)	Precursor ion (m/z)	Adduct ion
Feature 9	256.1491	27.6	257.1563	274.1827

Since the feature was found when reanalysing the data using different software, it provided an indication that the feature was in fact true and belonged to a metabolite.

Reproducible results with software are not only essential for verifying the trueness of a feature, but also (and even more importantly) measuring features repeatedly in different samples and on other platforms (alternative LC-QTOF). For this reason targeted analysis with LC-QTOF, on a pooled urine sample, was also part of verification and will be discussed in the next section.

3.4.1.2 Targeted analysis with LC-QTOF

The relative retention times calculated in Section 3.2.7 were used to set retention time windows for features. The software settings allowed the option of setting delta retention time windows, in this case three minutes, meaning that the window for a specific feature was six minutes long (three minutes before selected retention time and three minutes after). The windows were relatively wide but were essential in order to compensate for possible errors in the calculated retention time index. The initial collision energy used for collision induced dissociation was a moderate 10 V. Table 3.4 indicates the precursor ions, retention time, relative retention time and detected retention time of Feature 9.

Table 3.4: Targeted MS/MS analysis results obtained for Feature 9.

Feature number	Precursor ion detected (m/z)	Retention time (Venter *et al.,* 2015) (minutes)	Relative retention time (minutes)	Detected retention time (minutes)
9	257.1563 274.183	27.67	30.83	28.603

By determining relative retention times (Section 3.2.8), it was then possible to search for features in the retention time windows where they were expected to elute. A relatively large retention time window was allowed for each mass, despite the accuracy of the predicted retention times. This increased the chance of detecting the mass of interest in the respective regions. By allowing the quadruple to select the masses of interest, any false positive features from detector ringing (among others) were thus excluded.

Feature 9 was detected at 28.6 minutes with a mass of 257.151 as well as 274.15. Venter *et al*. (2015) reported 257.151 at a retention time of 27.67. The retention time shift was not as drastic as originally expected when studying the relative retention time, but this supported the decision to set up large retention time windows in the targeted MS/MS analysis (relative retention time is a predicted rather than a set retention time).

In conclusion, Feature 9 was verified on the first level of verification since the feature was detected during data reanalysis with a different software package (Progenesis QI). The feature was also detected in a different sample (pooled urine sample) on a different instrument (Agilent 1200 series LC system coupled to a 6510 Q-TOF mass analyser). From this it could be deduced that Feature 9 was not due to over-deconvolution or an analytical artefact, but could rather be attributed to a metabolite/fragment of a metabolite.

3.4.1.3 LC/MSD Trap XCT Plus analyser

The first aim in the analysis of fractions with the Ion trap MS was to detect the internal standards in order to predict the fraction that contained specific features. The second aim was to correctly identify internal standards, with fragmentation spectra using the METLIN database, to determine whether the selected process was adequate for characterizing the features through spectral matching (if such an entry existed in the selected database). Lastly, an attempt was made to characterize the features in the fractions.

The internal standards were found in fractions 4 (nor-leucine), 18 (2-acetamidphenol) and 24 (caffeine), respectively. Using the information provided by the internal standards, it was predicted that fraction 36 contained Feature 9. Figure 3.6A illustrates the fragmentation spectra of caffeine found in the METLIN database which indicates two distinct fragments for caffeine, with masses 138.0659 and 110.0715 when fragmentation is performed at a low energy. The ion trap data for fraction 24, containing the internal standard caffeine, is illustrated in Figure 3.6B.

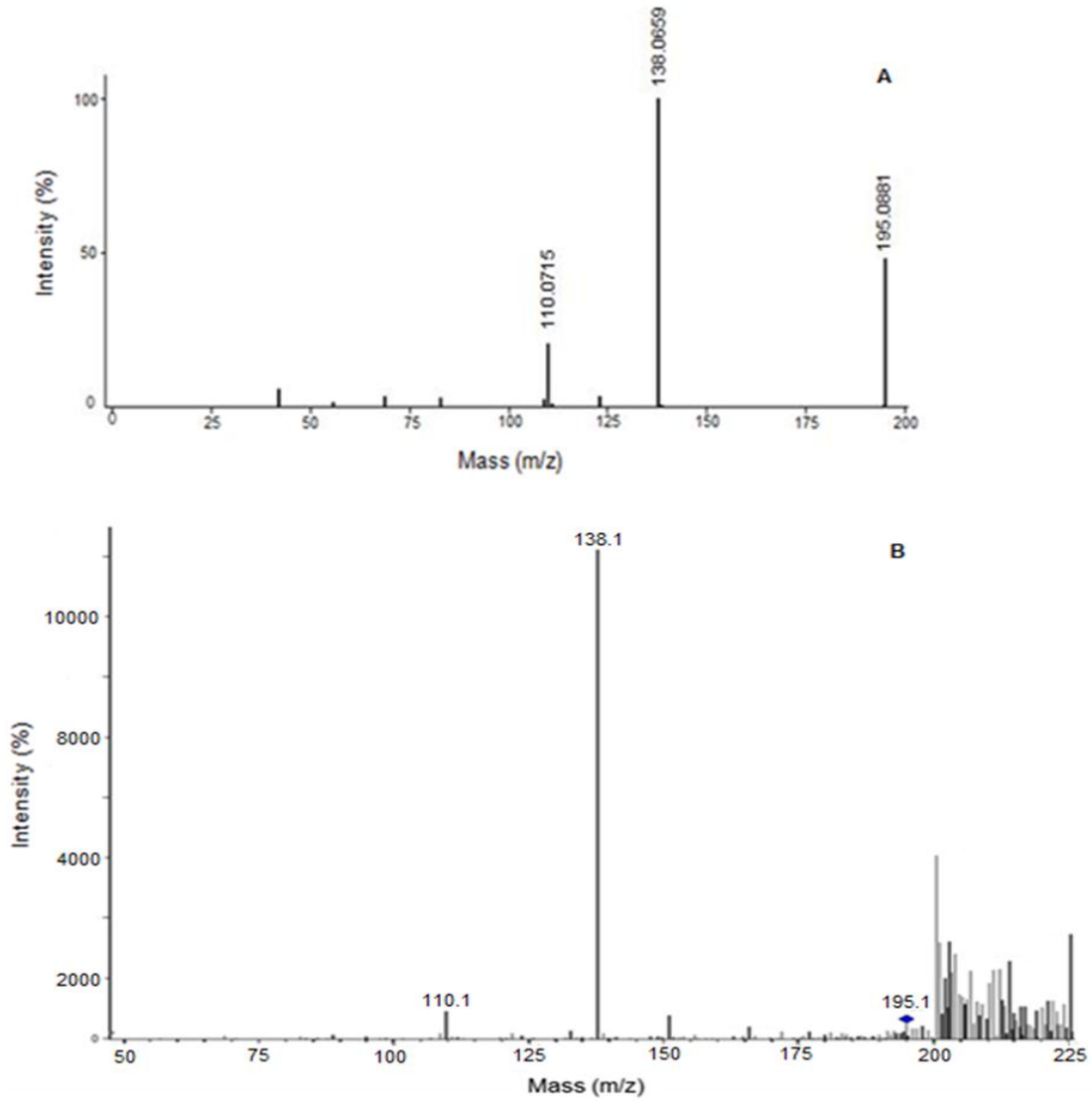

Figure 3.6: Fragmentation spectra of caffeine. Spectra of caffeine found in METLIN database (A) and isolated 195.1 mass in fraction 24 (B). The blue dot indicates the isolated 195.1 mass.

Despite significant background visible, it was possible to match the obtained spectra to that in the database. The same process was performed for the remaining two internal standards (results not shown).

With the correct identification of the internal standards, it was concluded that the selected process was adequate for characterizing the features through spectral matching.

In Figure 3.7, the fragmentation spectra of Feature 9 (obtained from analysis of fraction 35), indicated the two masses (257.0 m/z and 274.1 m/z) that had previously been linked to Feature 9 when reanalysing the raw data of Venter *et al*. (2015). The fact that the feature was found in fraction 35 supported the prediction of the fraction containing feature guided by retention time, meaning that based on retention time and detection of masses associated with the feature, it was safe to assume that Feature 9 was not an artefact of analysis but a true feature/metabolite.

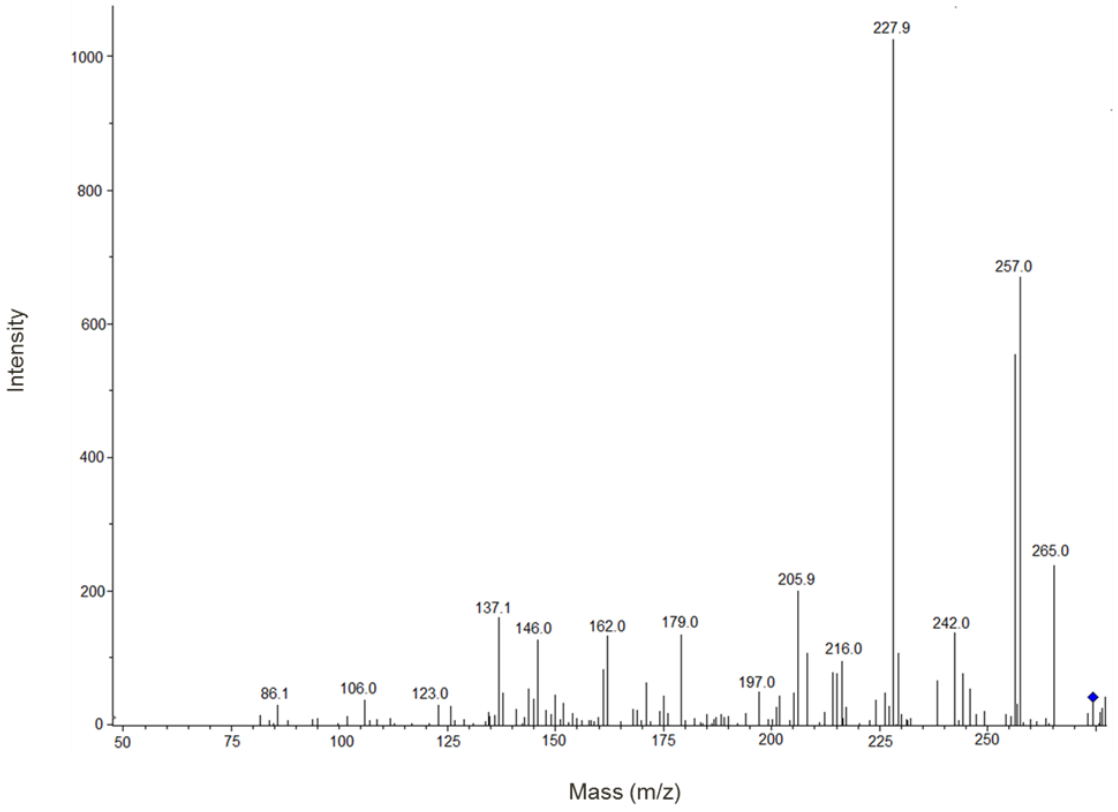

Figure 3.7: Fragmentation spectra of Feature 9 in fraction 35. The blue dot indicates the isolated 274.1 m/z.

Further characterization of the feature was impossible since clean fragmentation spectra could not be obtained, possibly because the feature concentration was too low, making it difficult to generate quality spectra. The intensity of each feature is crucial because the precursor ion intensity dictates the signal-to-noise ratio of the fragment ions (Yanes *et al*., 2011). Even with the internal standards, which had a high concentration (100 ppm), clean spectra (with no background) could not be obtained. However, the internal standards concentrations were high in the sample, thus generating better spectra than in the features.

Additionally the ion trap analyser lacks the time of flight (TOF) parameter, resulting in fragments without accurate mass. In order to match fragmentation spectra to spectra in a database, accurate mass measurements of precursor and product ions were required. For this reason it was decided to use the Synapt G2 Si hybrid IM-MS system which has certain capabilities to overcome some of the problems encountered when using the ion trap.

3.4.1.4 Synapt G2 Si hybrid IM-MS system

The pooled urine sample (described in Section 3.2.4) was analysed and fragmentation spectra for Feature 9 were obtained, as shown in Figure 3.8. In this figure, the low energy (A) and high energy (B) fragmentation spectra obtained. In Figure 3.8 (A), a single fragment of 274.1830 in the low energy channel (first channel) can be seen, with (B), the high energy channel (second channel), indicating two masses 257.1579 and 274.1886. These two masses correlate with the findings in Section 3.4.1.1 where Progenesis QI was used to reanalyse the raw data and findings in Section 3.4.1.3, and where fraction 35 (containing Feature 9) was analysed with an ion trap analyser. Additionally, Figure 3.9 shows extracted chromatograms for high energy channel (A), low energy channel (B) and the total ion chromatogram (C) for Feature 9, which indicate the elution time of 27.3 that correlated with the reported retention time described by Venter *et al*. (2015).

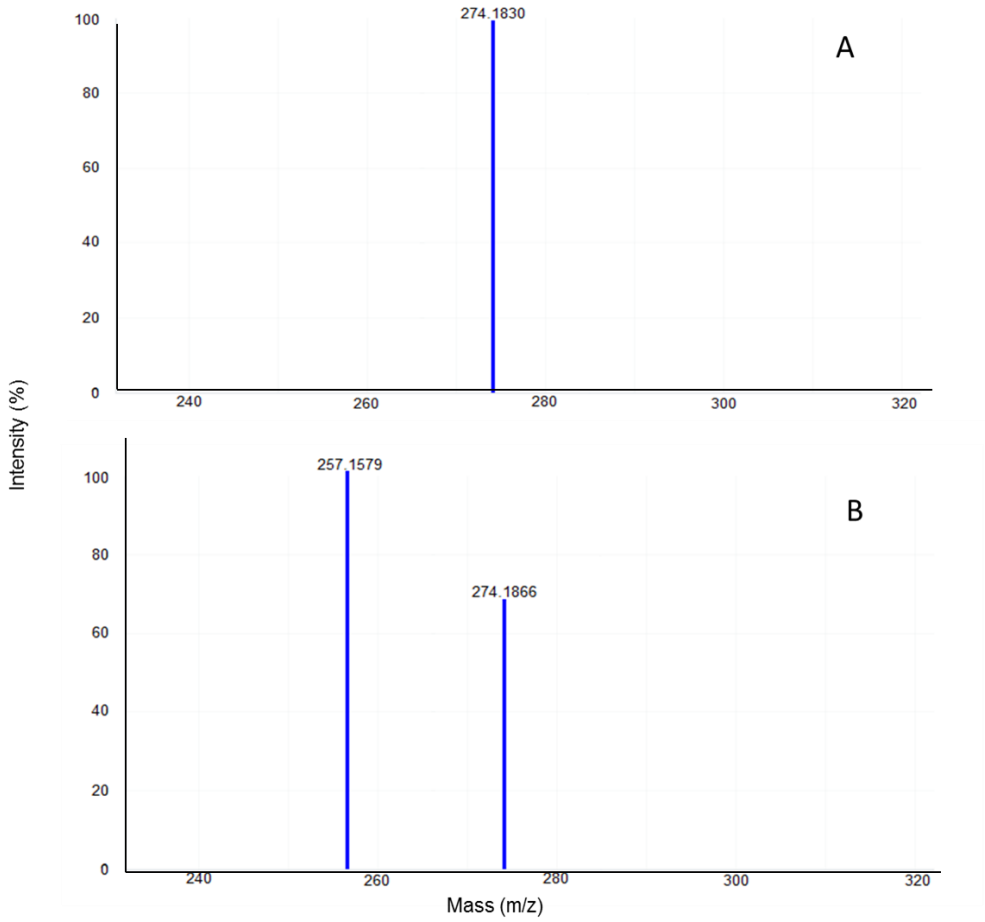

Figure 3.8: Fragmentation spectra of Feature 9. A) Chromatogram acquired in low-energy mode and (B) Chromatogram acquired in high-energy mode within the same experiment**.**

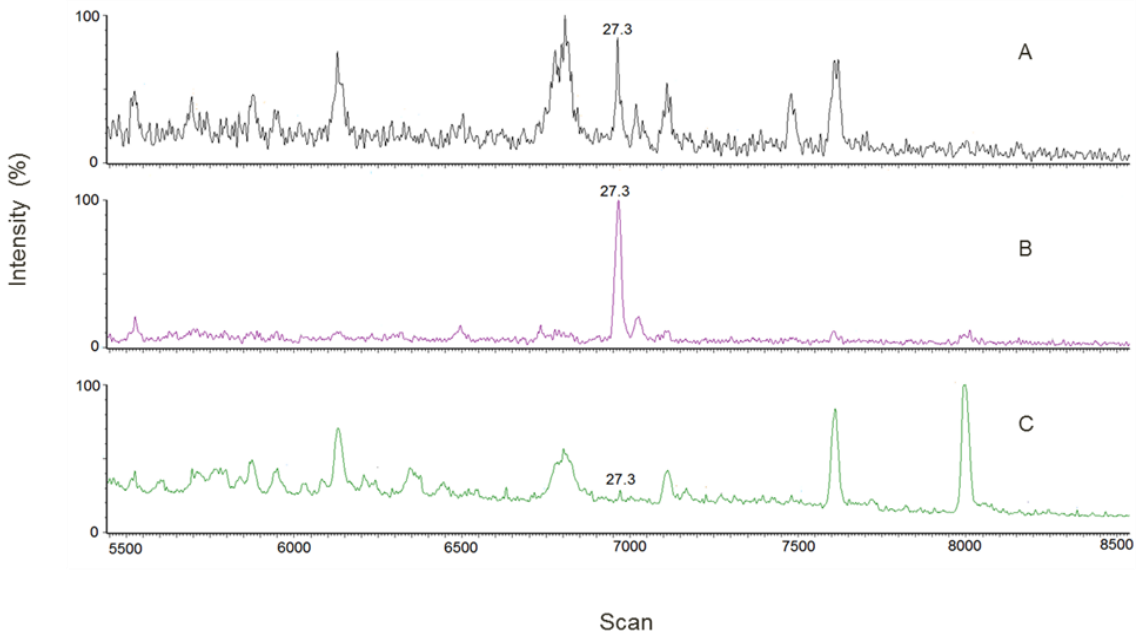

Figure 3.9: Representative chromatograms of Feature 9 in pooled urine sample using Synapt MS. Extracted chromatogram of feature 9 detected at retention time 27.3 in the first channel (low energy) (B) and second channel (high energy) (A). The total ion chromatogram (TIC) is shown in (C).

Even with the added advantages of Syntapt G2, sufficient spectra could not be generated to make a possible identification. The spectra in Figure 3.8 were exported to several databases (Metlin, MassBank) in an attempt to identify the responsible compound. Unfortunately, no identity could be found and it was decided to use *in silico* fragmentation as another way of identifying the compound. The need and advantage of *in silico* fragmentation is realized when the number of compounds with available experimental MS/MS spectra in all databases are considered. For example, Metlin has ~241 000 compound entries of which only ~13 000 have spectral information; HMDB has ~42 000 metabolite entries of which only ~6 000 have spectral information; and NIST have more than 200 000 compounds in its database of which only ~6 000 have MS/MS spectra.

The advantage of Progenesis QI over other software is that it can use Structure Data Format (SDF) files of several databases to create *in silico* MS/MS for identification purposes and to generate *in silico* fragmentation spectra for spectral matching. All features in the data set were matched to the accurate mass and *in silico* fragmentation spectra of molecular structures obtained from numerous databases CheBI (www.ebi.ac.uk/chebi), HMDB (www.hmdb.ca), Lipid maps (LMSD, http://www.lipidmaps.org) and Drug bank (DrugDB, http://www.drugbank.ca).

Table 3.5 represent the identification data for Feature 9 and include the compound ID, description, overall score and fragmentation score. The compound ID indicates the database in which the specific feature was found, in this case HMDB. An overall identity score is calculated for each possible identity using parameters such as accurate mass, retention time, ion mobility (collision cross-section) data, isotope spacing and fragmentation that match those in the database. The availability of parameters dictated the outcome of the score, for Feature 9 the score was 58 which are considered good when taking into consideration that from all the parameters only accurate mass, isotope spacing and fragmentation was available. A fragmentation score was also given, which indicated how well the theoretical fragments matched the measured fragments. In the case of Feature 9, a 100% match was obtained. Based on the identification confidence level, the findings of Feature 9 was at Level 2, with the next step requiring the matching of the retention time and mass spectra data to that of a reference standard, thus lifting the current level (Level 2) to Level 1 (the ultimate confidence level for metabolite identification, as described in Section 3.3.2). In doing so, a possible identification was made with HMDB for Feature 9. Table 3.5 shows the identification results of Feature 9, as obtained using Progenesis QI, when doing *in silico* fragmentation.

Table 3.5: Identification data of Feature 9 by Progenesis QI. The HMDB compound ID are given along with the description, formula, overall and fragmentation score as well as the ID level proposed by Schymanski *et al*. (2014).

Compound ID	Description	Formula	Overall score	Fragmentation score	ID level (Schymanski *et al*., 2014)
HMDB05935	Androstenol	$C_{19}H_{30}O$	58	100	2

The androstenol is a steroid compound not previously linked to RCDs. It should be noted that this is just a possible identification and must still be positively identified using a reference standard.

3.4.2 Verification of the remaining 11 features

The remaining features were verified with the procedure followed for the verification of Feature 9. Identical steps were followed and the results are summarized in Table 3.6.

Table 3.6: Verification results of the remaining features. The verification results of the remaining 11 features performed on the various instruments.

Feature number	Progenesis	LC-QTOF	Ion trap	Synapt
1	x	-	-	-
2	-	-	-	-
3	x	x	-	x
4	x	-	-	x
5	-	x	-	x
6	x	x	x	x
7	-	x	-	x
8	x	x	x	x
10	x	x	x	x
11	x	x	-	-
12	x	-	-	-

(-) Feature was not detected in the specific procedure. The (X) indicates that the feature was detected.

Deconvolution was reviewed to ensure Progenesis QI grouped all possible adducts to calculate an accurate neutral mass. A total of six of the nine features were detected with positive ionization and both of the two features of negative ionization were found with Progenesis QI. For Feature 1, one precursor ion and a possible adduct of 386.0256 $[M+K]^+$ were found. For Feature 3 adducts given were 547.8636 $[M+3H]^+$ and 821.2965 [M+2H]. Feature 3 was reported (Venter *et al*., 2015) with a mass of 1043.4677, in Progenensis QI the neutral mass was detected as 501.0894 with additional adducts of 524.0786 $[M+Na]^+$ and 540.0440 $[M+K]^+$ respectively. Feature 6 were reported as 314.1547, with Progenesis QI the adduct given was 332.1906 $[M+H+H_2O]^{+\cdot}$ For Feature 8 a mass of 222.1615 was reported Progenesis QI also gave adduct of 205.1566 $[M+H-H2O]^+$. After processing with Progenesis QI, features of negative ionization gave the following results: Feature 11 was detected in the charge state of 145.0215 $[M+H]^-$ and Feature 12 detected one precursor ion 175.0641 $[M+H]^-$ and a possible adduct of 155.0174 ((M-H2O)-H). Feature 2, 5 and 7 were not found in data reanalyzed with Progenesis QI.

Using targeted MS/MS analysis on the pooled urine sample, only seven (Features 3, 5-10) of the nine remaining features previously found in positive ionization mode were detected. For negative ionization mode only one (Feature 11) of the two features were detected. The remaining features (1, 2, 4 and 12) were not detected when using this targeted approach. One possibility is that the four remaining features had concentrations below the detection limit.

Since clean accurate mass fragmentation spectra could not be obtained with the ion trap mass spectrometer, it was difficult to determine which information was of value. In Table 3.6, a feature was only marked with an X when the findings could be correlated to accurate mass data from Progenesis QI where the adduct ions can be used and the accurate mass fragmentation data of the Synapt G2. Only Features 6, 8 and 10 could positively be linked to the information (nominal mass and retention time) already available for the specific feature. The fragmentation spectra obtained when isolating the 332.1 mass of Feature 6, visible is the 314.2 mass indicating a water loss. For Feature 8, the precursor ion of 222.2 was isolated and the 205.1 mass was visible, also indicating a water loss. These results correlate with the findings in the Progenesis QI data. For Feature 10, the fragmentation spectra of the ion trap and the Synapt could be matched to some extent. The data of the remaining features were inconclusive and may be due to the feature concentration being too low to obtain spectra that contained useful information. When the pooled urine sample was analysed with the Synapt G2, the following features were detected: Features 3, 5, 6, 7, 8 and 10. The accurate masses and retention times were used to search for the features.

A challenging part of the study was obtaining fragmentation spectra for features, in order to gather more information regarding each feature. Quality fragmentation spectra for some features were unobtainable and the reason for this was most likely due to the low concentration of the features. Even if quantity fragmentation spectra could be obtained, it did not necessarily mean that identification could be made using a database. For this reason, *in silico* fragmentation was particularly useful. Table 3.7 indicates the identifications made for features detected in the pooled urine sample using the Synapt G2 and data processing with Progenesis.

Table 3.7: Identification data of the remaining features by Progenesis QI.

Feature number	Compound ID	Description	Formula	Overall score	Fragmentation score	ID level (Schymanski *et al*., 2014)
5	HMDB38935	Kamahine C	$C_{14}H_{20}O_5$	37.9	0	4
6	HMDB32223	Diisopentyl thiomalate	$C_{14}H_{26}O_4S$	51.8	72.3	3
8	HMBD 36790	Kobusone	$C_{14}H_{22}O_2$	38.2	0	4

In Table 3.7 only one feature had a fragmentation score, which means for this feature alone, a match was found when doing *in silico* fragmentation. The remaining features were identified by matching their accurate mass to a database. The features that are not listed in Table 3.7 could not be identified. Unlike in the case with Feature 9 that had a fragmentation score of 100%, the identifications made for the three features listed in Table 3.7 is just possible identification (no or low fragmentation score) requiring further investigations and positive identification to link them to any biological system.

Additionally, data extraction programs used to generate spectra was Masslynx and MSE viewer along with Progenesis QI and various databases was used to make a potential identification. No definite identification could be made. Thus, due to the uncertainty of the true identifications of the features, the original annotated names were used when referring to verified features that form part of the improved biosignature.

3.5 Final feature selection

In order to improve the biosignature, the features that could not be detected a second time around, were eliminated. In doing so, the features that were actual features of interest and could be measured and detected repeatedly, were included in the improved biosignature. For the elimination process, criteria were established in order to decide which of the features would be eliminated. The criteria stated that: 1) the features had to be found in the original raw data from Venter *et al*. (2015) when reprocessed by Progenesis QI, because this data was the original data used to construct the biosignature; and 2) The feature had to be detected in at least two of the alternative platforms. The selected features are shown in Table 3.8. These features will form part of the improved biosignature.

Table 3.8: Features obtained by LC-QTOF analysis that will form part of the improved biosignature.

Annotated name	Feature number	Mass	Retention time
C23HNO8S4	3	547.8636	12.7
C16H26O4S	6	314.1547	24.9
C14H22O2	8	223.1688	24.7
Androstenol	9	257.1563	27.67
Tetrapeptide	10	444.186	21.7

3.6 Discriminative power of verified features

The discriminative power of the verified features was evaluated using a PCA score plot. The data used for this discriminative power evaluation were the data matrix of Venter *et al.* (2015). This gave an indication of the ability of these features to distinguish between the RCD patient group and the CRC group.

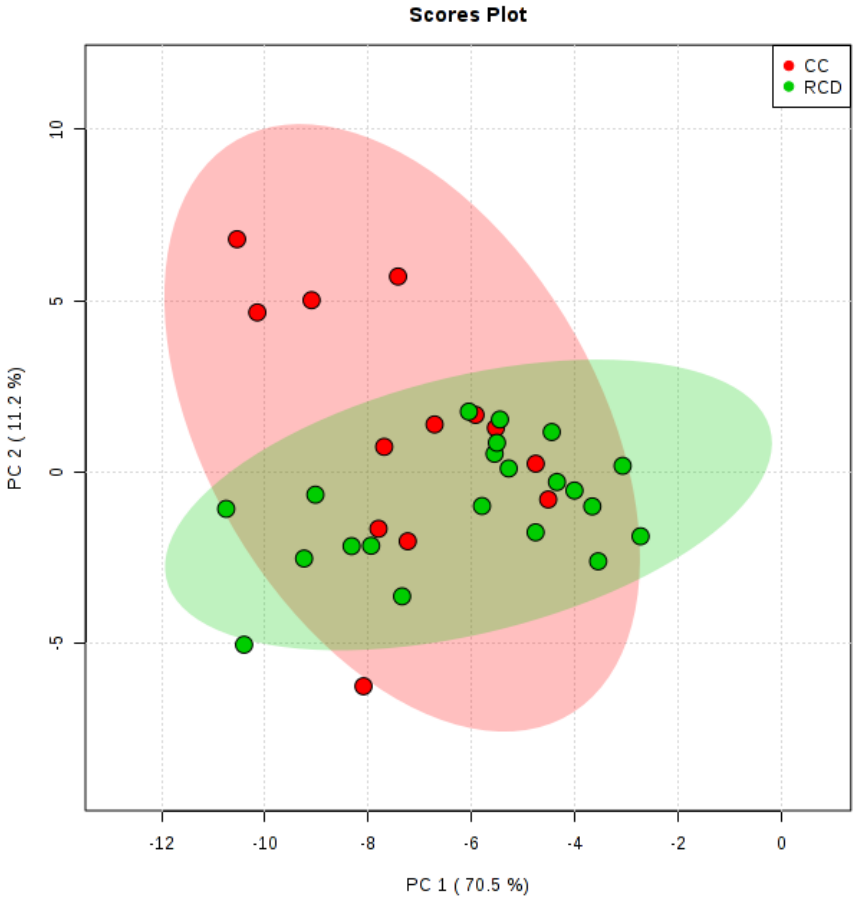

Figure 3.10: PCA score plot of the verified features. The figure displays the discriminative power of the RCD patient group (green) and CRC group (red) with the use of the verified features.

The PCA in Figure 3.10 was constructed using the verified features (listed in Table 3.8). In Figure 3.10, the RCD group display clustering to some extent, while the CRC group are more scattered across the plot (with some of the CRC samples being completely separated from the others). It was to be expected that the RCD group would not cluster together tightly due to the diversity of clinical symptoms found within the RCD group and the heterogeneity of RCDs. The CRC samples that were scattered among the RCD patients suggested that based on the verified features alone, the metabolite profiles of these patients were similar. By excluding the seven remaining features from the biosignature, some of the initial discriminative power was lost, possibly due to covariance of the 12 features of the proposed biosignature. Therefore, the verified features were unable to exploit the biological variance between these two groups.

3.7 Summary

The aim of this part of this study was firstly to verify that the features could be detected when using different software, and secondly, that the features could be detected when using a different instrument. The ultimate aim was to eliminate those features that could not be detected in the verification process in order to improve the biosignature proposed by Venter *et al*. (2015). Lastly, characterization was also attempted. Five of the 12 features that were included in the proposed biosignature could be verified. The remaining seven were eliminated by the criteria used for elimination. The characterization of features proved the most challenging part of the study with only four potential identifications.

The discriminative power of the five verified features was evaluated. Results indicated that when using the five verified features alone, the initial discriminative power that the proposed biosignature had was lost. The proposed biosignature was able to partially differentiate between these two groups, but the ultimate goal would be to obtain complete separation between these groups. However, complete separation is not necessarily needed nor is it always possible, when considering that some of the CRC patients had enzyme activity scores very close to RCD positive scores. In order to improve the biosignature, additional platforms with the ability to expand metabolome coverage need to be included. Analyses using the same cohort could yield additional metabolites that could be included in the improved biosignature, thus increasing the discriminative power.

CHAPTER 4

GAS-CHROMATOGRAPHY MASS SPECTROMETRY ANALYSIS OF PATIENT SAMPLES

4

4.1 Introduction

The biosignature proposed by Venter *et al*. (2015) was able to partially differentiate between RCD and CRC samples that presented with similar symptoms. However, since only untargeted LC-MS analysis was performed in that study, only a section of the metabolome was analysed and considered for the biosignature. No single platform is able to measure all metabolites present in a biological sample, thus using multiple analytical platforms in metabolomic studies can expand metabolome coverage (Dunn *et al*., 2011). It was decided to include untargeted GC-MS analyses that could expand this biosignature in terms of specificity and sensitivity (Gullberg *et al*., 2004). GC-MS is a popular technique for metabolomic studies for a number of reasons, including high separating power and reproducibility as well as having electron impact libraries that assist in metabolite identification (Koek *et al*., 2006).

In this chapter, all aspects for the expansion of the biosignature using an untargeted GC-MS method, will be given and discussed. As a starting point, method standardization was performed. Next, the method was implemented on a sample set of CRC and RCD patients (Section 3.2.3). Statistically significant discriminatory features were identified with uni- and multivariate statistical methods, in an attempt to expand the RCD biosignature.

4.2 Materials and methods

4.2.1 Reagents and chemicals

Acetonitrile (Honeywell, Burdick and Jackson, cat # BJ015CS), hexane (Sigma-Aldrich, cat # 296090), methoxyamine hydrochloride (Aldrich, cat # 226904), O-bis(trimethylsilyl) trifluoro acetamide (BSTFA) with trimethylchlorosilane (TMCS) (Fluka, cat # 33155), pyridine (Sigma-aldrich, cat # 270970) and water (Honeywell, Burdick & Jackson, cat # BJ365CS) were general reagents used throughout sample preparation.

Nor-leucine (Sigma-Aldrich, cat # 74560), 2-acetamidophenol (Sigma-Aldrich, cat # A700025GA) and 3-phenylbutyric acid (Aldrich, cat # 116807) were used as internal standards. For the extraction of organic acids ethyl acetate (cat # 141-78-6) and diethyl ether (cat # 60-29-7) were used and were purchased from Merck chemicals.

4.2.1.1 Internal standard preparation

Each of the three internal standards (nor-leucine, 2-acetamidophenol and 3-phenylbutyric acid) were prepared as stock solutions with a concentration of 1000 ppm. These 1000 ppm stock solutions were diluted to a final concentration of 500 ppm. A volume of 10 µL of each internal standard (500 ppm) was added to each sample, and following the preparation steps, the sample was dried and dissolved in a total of 100 µL derivitization reagents, resulting in a final concentration of 50 ppm for each internal standard.

4.2.1.2 Oximation and silylation

Oximation reagent was prepared by weighing 200 mg of methoxyamine hydrochloride in a Kimax tube and dissolving it in 10 mL of pyridine. After the samples were dried under N_2 gas, 50 µL of the prepared oximation reagent was added and then vortexed for 1 minute to dissolve the dried compounds. The samples were incubated for 60 minutes at 60ºC. Silylation was performed by adding 50 µL of BSTFA with 1% TMCS to the oximated samples and incubated again for 60 minutes at 60ºC.

4.2.2 Instrumentation

The GC-TOF-MS-system consisted of an Agilent 7890 GC system coupled with a Leco Pegasus HT mass analyser, including an Agilent 7693 auto sampler. Chromatography was performed on a Restek RXi® - 5Sil-MS column (30 m X 250 µm X 0.25 µm). Of each sample, 1 µL was injected per run with 1:10 split. The front inlet temperature was kept at 250°C throughout the entire run. The initial oven temperature was held at 70°C for 1 minute after injection, followed by an increase of 15°C per minute to 250°C. The oven temperature was then increased at a rate of 18°C/min, where it was maintained for 1 minute equating to a total run time of ~16.7 min per sample. Hydrogen was used as carrier gas at a constant flow of 2.5 mL/min. The transfer line was maintained at 225°C and the ion source temperature at 200°C, for the entire run. Acquisition was delayed for the first 135 seconds, which served as a solvent delay. Data was captured with an acquisition rate of 20 spectra (50 – 950 m/z) per second.

4.2.3 Sample preparation

4.2.3.1 Organic acid extraction

Metabolome coverage and repeatability of the organic acid extraction method described by Reinecke *et al*. (2012) was evaluated. Briefly, urine was transferred to Kimax tubes and acidified using 230 µL of 5 M HCl. Of ethyl acetate, 3 mL was added to each sample and the mixture was shaken on a rotary wheel for 30 minutes. The mixture was centrifuged for 10 minutes at 1300 x *g* and the upper ethyl acetate phase was transferred to a clean Kimax tube. Of diethyl ether, 3 mL was added to the water phase and shaken for a further 15 minutes. After centrifugation (1300 x *g*, 10 minutes), the upper phase was removed and added to the ethyl acetate. A spatula of sodium sulphate powder was added to the ethyl acetate/diethyl ether mixture to remove any residual water. After a subsequent centrifugation step, the organic phase was transferred into a vial and dried under nitrogen at 40°C. Dried samples underwent oximation and silylation (described in Section 4.2.1.2). The derivatised sample was transferred to a vial for GC-MS analysis.

4.2.3.2 Deproteinised urine assay

Minimal sample preparation was employed for the untargeted analysis. The predetermined volume of urine, internal standard and ice cold acetonitrile (three times the volume of urine) were added to a microcentrifuge tube. Each sample was centrifuged at 25 000 x *g* for 10 minutes. The supernatant was transferred into a vial and evaporated to dryness under nitrogen at 40°C. Once the samples were dried, oximation and silylation (described in Section 4.2.1.2) were performed, and the samples were transferred to a new vial for analysis.

4.2.3.3 Quality control

Quality control (QC) samples are used in metabolomics experiments to evaluate and compare the quality and performance of a particular analytical process. QCs are therefore crucial for ensuring that the acquired data is suitable for data analysis and produces robust as well as valid data sets from which biological conclusions can be made. QCs are also used to determine and correct for any inconsistencies that may occur during extraction or analysis. These QC samples were prepared by combining equal amounts of each sample in the entire RCD patient sample cohort. The QC sample creatinine value was determined (4,539 mmol/L) and sample preparation was identical to that of the experimental groups (described in Section 4.2.3.2).

4.2.4 Data analysis

4.2.4.1 Data extraction

The raw data obtained from the analyses on the biological samples were in the form of several chromatograms and mass spectra. GC-MS data consists of retention times, molecular masses and peak intensities (Dettmer *et al.,* 2007). The ultimate goal of data extraction is to convert chromatograms and mass spectral data into a universal format so that mathematical and statistical analyses can be performed. In the process of data extraction, there are three main steps involved that include peak picking, deconvolution and alignment (Luedemann *et al*., 2008).

Leco Corporation ChromaTOF software was primarily used for peak detection and mass spectral deconvolution. Deconvolution is used to separate co-eluting peaks and goes hand-in-hand with peak detection. Ultimately, deconvolution is used to resolve complicated mass spectra of co-eluting compounds into individual mass spectra (with the correct intensity) for each detected analyte. This allows rapid and reliable quantification of co-eluting compounds in order to extract valuable information (Lisec *et al*., 2006).

The final step in data extraction is peak, feature or compound alignment. The main aim of peak alignment is to align peaks that have identical mass spectra and corresponding retention times across all samples. In this study, Statistical compare, a function of Leco Corporation ChromaTOF software package, was used to align the peaks of all urine samples that were analysed.

4.2.4.2 Data pre-processing

Pre-processing of metabolomics data can be described as the final editing of data to the point where it is ready for statistical analysis (Liland, 2011). The steps taken in pre-processing of the data in this study, included: zero filtering, missing value replacement and CV filter. The data was also evaluated to detect batch effects and possible outliers.

Firstly, zero filtering was used to remove non-biological features, or those features with no statistical value, from the dataset (effectively creating a more complete data set). In this study, the features with >40% missing values, in any of the experimental groups, were removed from the dataset (Lindeque *et al*., 2013) as part of the zero filtering criteria.

Missing values are a common occurrence in datasets of mass spectrometry experiments in metabolomics studies. These missing values can originate from multiple sources, including biological and technical sources (Hrydziuszko and Viant, 2012). In many cases, these missing values are a result of certain compounds being below the instrument's detection limit (Xia *et al*., 2009). In this study, the missing values were replaced with half of the detection limit (50% of the minimum value found in the dataset).

Data normalization is used to reduce the influence of experimental factors. Normalization with an internal standard eliminates any variation that could occur, in order to make samples more comparable. Normalization of the peak areas relative to the internal standard (3-phenylbutyric acid) was performed by dividing the peak areas with the internal standard peak area.

The CV filter allows a cleaner data matrix that removes compounds with large analytical variation. As a quality assurance step, QC samples are analysed over time and provide a reliable indication of features that should be removed from the dataset. In this study, a CV value of 50% was chosen as the cut-off. Thus, all features in the QC samples with a CV value >50% were removed from the dataset.

The term batch effect can be described as the systematic non-biological differences between batches of samples due to various factors that occur during sample handling or during analysis. A batch effect can occur in different batches (inter-batch) or in the same batch (intra-batch) (Luo *et al.,* 2010). Due to the relatively short run time per sample, it was decided to analyse all the samples in a single batch. This single batch analysis excludes batch-to-batch variation. As a quality overview, the dataset was visually evaluated for any intra-batch effects. No visible effect was seen and for this reason no batch-correction methods were applied.
Lastly, a heat map was used to visually inspect the quality of data. A heat map can be described as the relative abundance of compounds detected in samples, represented by colour intensity (Ivanisevic *et al*., 2014). Hierarchical clustering begins by considering each sample as a separate cluster, followed by combining them until all the samples belong to a single cluster, thus generating a heat map (Xia *et al*., 2009).

4.2.5 Method standardization

For repeatable and reliable analyses, method standardization was performed. Sample preparation is an essential step that serves the following purposes: 1) extraction of analytes from any given biological matrices; 2) conversion of analytes to a format that is suitable and

compatible with a specific analytical platform; and 3) removal of matrix components that may interfere in the analysis (Fernández-Peralbo and Castro, 2012). In untargeted metabolomics, sample preparation poses a challenge since there is no single method that has the ability to extract all metabolites (from various metabolites classes) within a sample (Álvarez-Sánchez *et al*., 2010b).

Since sample preparation is essential, two different metabolite extraction methods were considered. An organic acid extraction assay was compared to a minimal sample preparation method, namely the deproteinised urine assay. Repeatability studies were carried out by preparing and analysing the same sample multiple times. The different extraction procedures were carried out five times per procedure on identical samples. Additionally, it was decided to compare the effect of organic acid extraction to the minimal sample preparation method in order to determine whether organic acids previously linked to RCDs, could be detected in the samples using different sample preparation methods.

4.2.5.1 Sample preparation and analyses

A pooled urine sample was prepared (described in Section 3.2.4) and the creatinine concentration was determined to be 4.539 µmol/L. Two separate sample preparation methods were followed. For the first method, minimal sample preparation was performed. A volume of 55 µL urine (representing 0.25 µmoles of creatinine), 10 µL of each internal standard and 165 µL acetonitrile were placed in an microcentrifuge tube and prepared as described in Section 4.2.3.2. For the second method, organic acid analyses were performed and entailed the mixing of 55 µL urine, 10 µL of each internal standard and 360 µL water, followed by the preparation described in Section 4.2.3.1. Each of the two extraction methods were repeated 5 times on identical urine samples. Figure 4.1 provides a summary of the sample preparation steps that were followed.

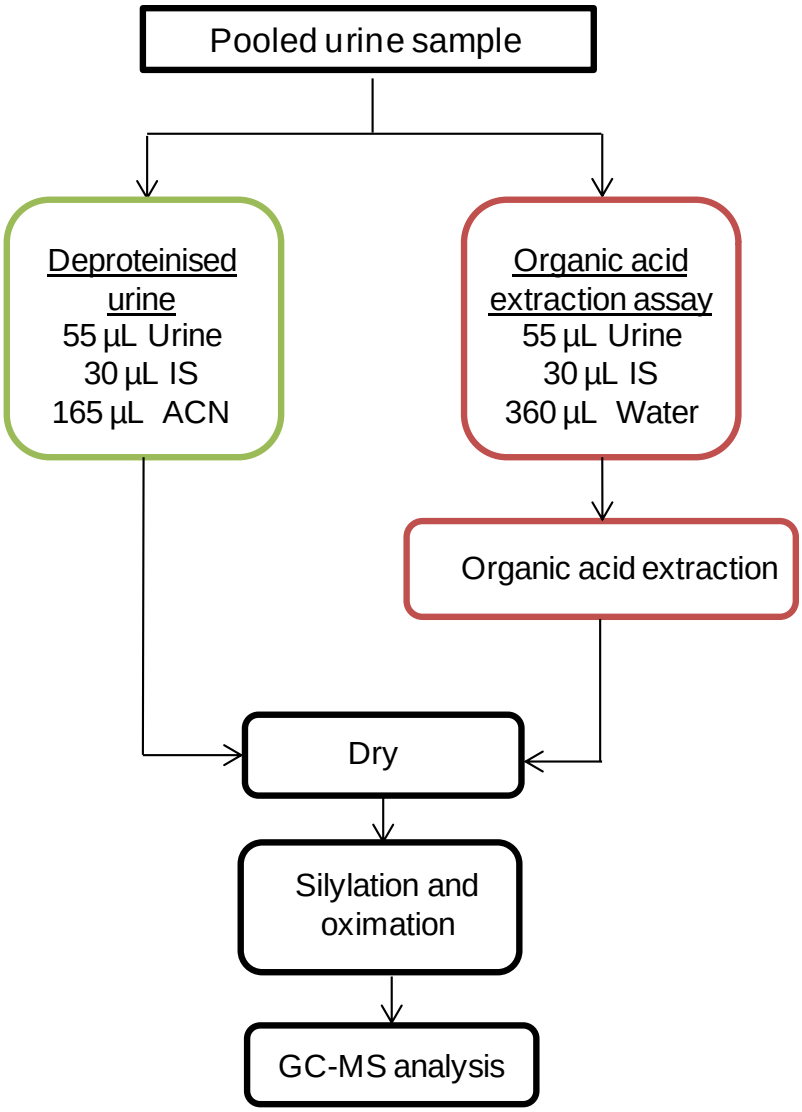

Figure 4.1: Workflow of the different sample preparation methods. Urine samples were prepared with two different methods: (1) deproteinised with acetonitrile and (2) organic acid extraction. IS: internal standard, ACN: acetonitrile

4.2.6 Biosignature expansion

4.2.6.1 Biological samples

Ethical approval, patient selection and experimental groups were described in Section 3.2.3. The urine samples (Section 4.2.5.1) were analysed according to the procedures described in Section 4.2.3.2.

The following run order was used:

QC1, QC2, QC3, Samples 1-7, QC4, Samples 8-14, QC5, QC6, Samples 15-21, QC7, Samples 22-28, QC8, QC9; Samples 29-35, QC10, Samples 36-41, QC11, QC12

In order to ensure that quality data was generated, QC samples were added throughout the batch. As can be seen in the above run order, the batch was initiated by the analysis of three QC samples, in order to condition the column. In the remainder of the batch, a QC sample was added after every seventh sample. The addition of two QC samples at certain positions was to validate batch correction, if necessary.

4.2.6.2 Statistical analysis

Following the data pre-processing described in Section 4.2.4.2, log transformation was performed prior to any statistical analyses. Log transformation firstly ensures normal distribution of the data and secondly limits the effect of more abundant compounds when performing multivariate analyses, thus ensuring that less abundant compounds also contribute to the grouping (van der Berg *et al.,* 2006). Following log transformation, the dataset was ready for statistical analysis. Not only is the generation of high quality data crucial in metabolomics studies, but also the strategy that is chosen for analysing the data (Madsen *et al.,* 2009).

Statistical analyses were performed in three phases: a data overview phase; a feature selection phase; and finally, the biosignature discriminative power testing phase. The goal of the first phase was to get an overview of the data by including the entire data matrix that was obtained from various steps that were performed prior to statistical analysis. The statistical methods used in this phase included a PCA plot and multivariate ROC modelling (using the support vector machine as the underlying statistical algorithm). The second phase was the feature selection phase where various univariate and multivariate statistical tests were used to identify a small number of important features that could potentially distinguish between the experimental groups. In the third phase, the discriminative power of the top ranked features was tested.

Biological variances that may be relevant for understanding disease or diagnosis are often masked by irrelevant variance, such as diet, ethnicity and gender. For this reason, it is advantageous to use multiple statistical tests to highlight the important features that may be able to differentiate between disease and controls (in this case, RCDs and CRC).

The important features identified using the univariate and multivariate statistical tests were listed, and the average rank of each feature across all the tests was determined. The top features that performed the best overall, based on a ranking plot, were then selected to expand the biosignature. In Figure 4.2, a summary of all statistical tests performed for feature selection, is illustrated.

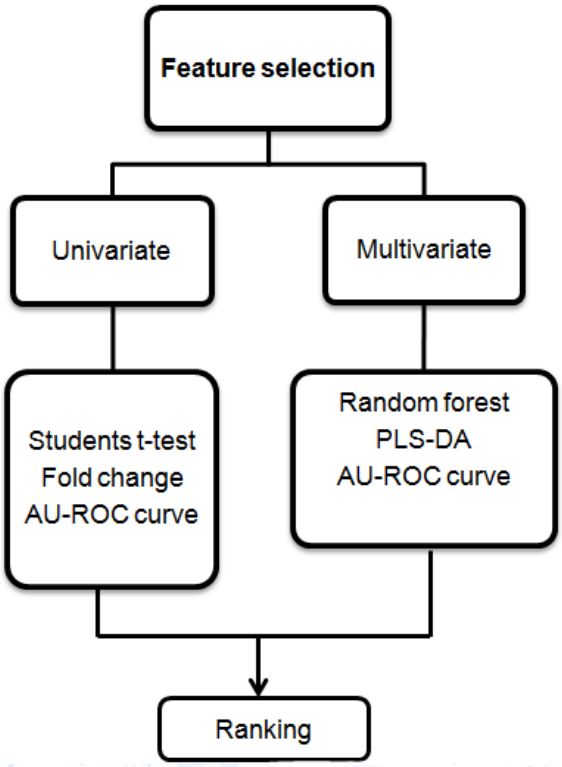

Figure 4.2: Summary of statistical analyses performed for feature selection. Univariate and multivariate methods were used in statistical analysis feature selection.

The following univariate, bivariate and multivariate methods were used when conducting statistical analyses on the data matrix obtained using both ESI methods. The univariate tests used for feature selection consisted of a Student's *t*-test, fold change and area under the receiver-operator curve (AU-ROC). The multivariate tests used for feature selection consisted of partial-least squares discriminant analysis (PLS-DA), random forest and bivariate AU-ROC. With the latter test, the ROC used when two features are combined, is shown.

The aim of a *t*-test is to determine whether the means of two groups are different. Once a *t*-value is established, a *p*-value is calculated that could be used to determine whether or not this difference is statistically significant. A small *p*-value of less than 0.05 is considered statistically significant (Xia *et al*., 2009).

Fold change can be described as the magnitude of the difference between two groups. Features of relevance within a dataset can be identified by ranking these features according to the fold change calculated (Vinaixa *et al.,* 2012). A fold change greater than one was considered statistically significant.

An AU-ROC curve can be seen as a bi-dimensional graph that illustrates a classifiers performance and a summary of its diagnostic performance. The interpretation of area under the curve (AUC) can be described as the probability that the classifier will rank a randomly chosen positive sample higher than a randomly chosen negative sample. When all of the positive samples are ranked before the negative samples, the numerical value of the AUC equals 1.0. When the AUC equals 0.5, it means that the classifier randomly classifies subjects as either negative or positive (Xia *et al.,* 2012). Features (univariate ROC), or combinations thereof (bivariate ROC), with an AUC value >0.75 were regarded as significant.

PLS-DA is a supervised method which utilizes various linear regression techniques to determine the direction of maximum covariance between the given data set and class membership. These analyses have the ability to perform classification as well as feature selection. Cross validation is used to choose the optimal number of components in order to establish classification. PLS-DA is used to identify compounds that contribute the most to separation between two sample groups. This is achieved by the ranking of compounds according to the variable important in projection (VIP), which can be explained as the weighted sum of squares of the partial-least squares loadings (Xia *et al*., 2006). Features with a PLS-DA VIP score > 1 were regarded as significant.

Random forest can be described as a combination of decision trees (also referred to as classification trees). These decision trees are constructed by bootstrap sampling, which creates multiple sets of randomly chosen compounds. In effect, these decision trees result in a powerful classification method that can handle a large number of variables and that can also measure the importance of a variable (Liland, 2011). Therefore, random forest merely ranks the variables according to their inclusion in the multiple discriminatory models tested. Since this test does not contain any clear cut-off value, it was decided to regard the top 20 variables as significant. The test settings included the growth of 1000 forests (subsets) using 10 features at a time (per subset).

Following the multiple multivariate and univariate statistical tests performed for feature selection, the important features were identified and the average rank of each feature was determined based on the performance of that feature. The univariate tests highlight discriminative features that differ significantly between the groups, when considered alone. However, more often than not, it is the ratio/combination of two or more features that can distinguish between the groups, which the multivariate tests must highlight. The latter is especially important when the mitochondrion is perturbed, as several metabolites are usually affected.

4.3 Results and discussion

4.3.1 Method standardization

The following results were obtained by evaluating the two different sample preparation methods (organic extraction method vs. minimum sample preparation method). The methods were evaluated in terms of repeatability and organic acid coverage using CV distribution, PCA plots and targeting a list of 25 organic acids previously associated with RCDs.

The distribution of the CV, for the detected compounds, was used to compare the repeatability of the two different extraction methods. PCA was used to establish whether or not a natural grouping existed between the sample groups based on their metabolite profiles (Abdi and Williams, 2010).

4.3.1.1 Method Variance

The repeatability of the two extraction methods were compared using the distribution of the calculated CV values. CV values were calculated using the relative, normalized concentrations of all compounds that were detected following GC-MS analysis of the extracted QC sample repeats.

The CV equals the standard deviation divided by the mean and is expressed as a percentage, as seen in Equation 4.1. The mean and standard deviation values for the deproteinised urine and organic acid extraction were calculated using the abundances of the features detected in the 10 samples.

Equation 4.1: Calculating the coefficient of variation (% CV)

$$\% \text{ CV} = \frac{\text{Standard deviation}}{\text{Mean}} \times 100$$

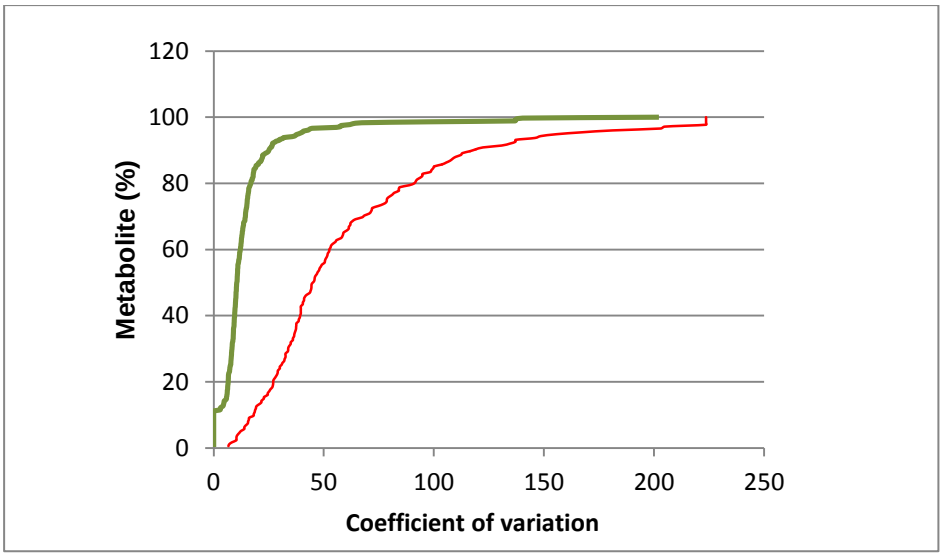

Figure 4.3: Deproteinised urine and organic acid extraction CV distribution. The green line represents the deproteinised urine and the red line represents the organic acid extraction method.

Figure 4.3 represents the CV distribution of the features found using the two different methods, organic acid extraction method (red line) and deproteinised urine method (green line). Table 4.1 provides a summary of the CV values for each extraction method.

Table 4.1: Summary of the CV values obtained by comparing the organic extraction method to the deproteinised urine method.

CV range	Organic acid extraction	Deproteinised urine
<10	2%	45%
10-20%	11%	41%
20-30%	11%	8%
30-40%	19%	2%
40-50%	12%	1%
>50%	45%	3%

According to the Food and Drug administration (FDA), a CV value of 15-20% is acceptable for targeted analyses, but when dealing with untargeted analyses, more variation is a common occurrence (t'Kindt *et al.,* 2009). When analysing the CV distribution data obtained for the two methods, it was clear that the samples that were deproteinised were more suitable for reliable GC-MS analyses, where features of importance could be more accurately quantified and identified in order to compile a biosignature. A greater percentage of the features had a CV value within the range of 0-10% for the deproteinised urine, thus indicating that the repeatability was acceptable.

PCA involves mathematical procedures that transform a number of possible related variables into smaller numbers of unrelated variables, called *principal components* (Saccenti *et al*., 2013). This is an unsupervised multivariate method that aims to display maximum variance between two groups. The PCA score plot is also useful for showing overall variance in the two sample groups.

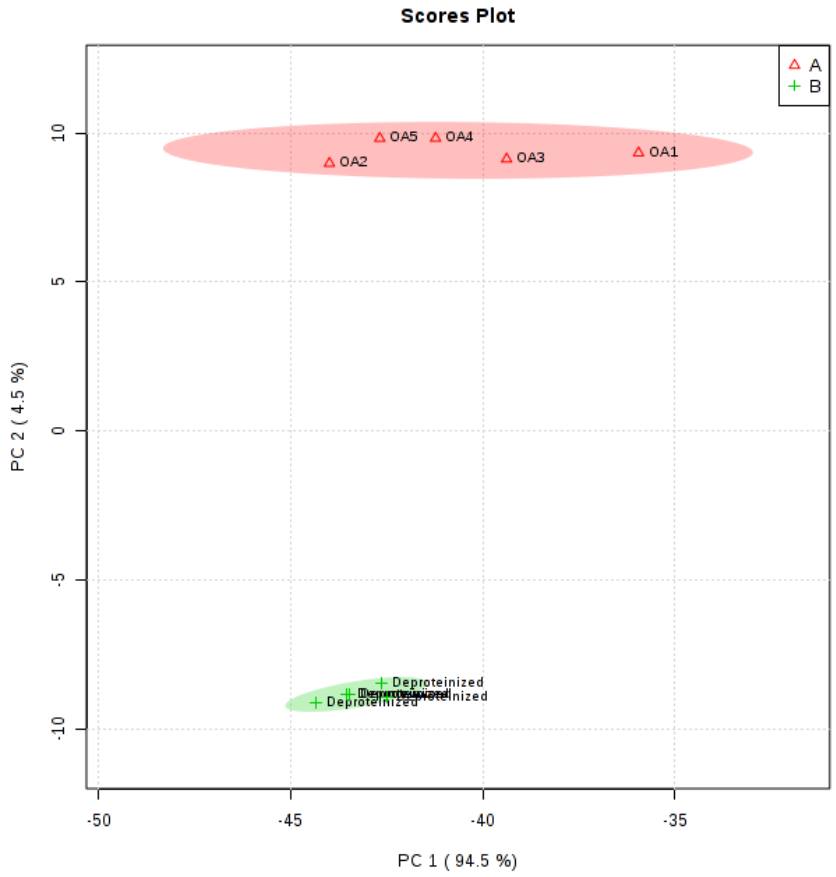

Figure 4.4: PCA score plots of different sample extraction methods. Two dimensional PCA score plot of the five samples prepared with the organic acid extraction assay (group A in red) and five samples prepared with the deproteinised urine assay (group B in green).

The PCA score plot in Figure 4.4 illustrates the natural grouping of group A and group B. Group A represents the samples prepared by the organic acid extraction while group B represents the deproteinised urine samples. Clear separation of the organic acid extraction samples (at the top in red) and deproteinised urine assay samples (at the bottom in green) could be seen. The repeatability of the deproteinised sample group exceeded the repeatability of the organic acid extraction group, as seen by the closer clustering of the deproteinised sample group. The organic acid extraction group lacked precision, as could be seen by the wide distribution of the red plots in the PCA. PC 1 showed the largest variation in the data and in the above example, it could be seen that 94.5 % of the inter-group variance was accounted for.

4.3.1.2 Metabolome/organic acid coverage

By comparing the compounds detected using the two sample preparation methods; it was possible to evaluate the quality of the method that was used to detect important metabolites. To do this, the organic acid metabolites previously associated with RCDs were compared. The organic acids listed in Table 4.2 were previously (but not exclusively) linked to RCDs. Reinecke *et al*. (2012) and Smuts *et al.* (2013) used the same South African RCD cohort and identified a list of organic acids associated with the disease. Reinecke *et al*. (2012) disclosed 24 metabolites in urine that were previously linked to RCDs, as well as numerous other organic acids that had not previously been linked to RCDs. Smuts *et al*. (2013) proposed a urinary biosignature that consisted of 13 metabolites, 6 of which were organic acids. Five of these organic acids were included in the study performed by Reinecke *et al*. (2012). The study conducted by Smuts *et al.* (2013) added one additional organic acid to the list, namely 3-hydroxyvaleric acid. Since the same South African RCD cohort was used in the current study (but with a different control group), focus was placed on the organic acids displayed in Table 4.2.

Table 4.2: Specific metabolites used to compare the two different sample preparation methods.

Name	Deproteinised urine assay	Organic acid extraction assay
1. **2-keto-ocatanoic acid**	n.d	n.d
2. **2-methyl-3-hydroxybutyric acid**	X	X
3. **3-hydroxyadipic acid**	X	n.d
4. **3-hydroxyisovaleric acid**	X	X
5. **3-hydroxysebacic acid**	n.d	n.d
6. **3-methoxy-4-hydroxyphenylacetic acid**	X	X
7. **3-methyladipic acid**	n.d	n.d
8. **3-methylglutanonic acid**	n.d	n.d
9. **4-Hydroxyphenylacetic acid**	X	X
10. **Ethylhydracrylic acid**	n.d	n.d
11. **Fumaric acid**	X	X
12. **Glutaric acid**	X	X
13. **Glycerol**	X	n.d
14. **Malic acid**	X	X
15. **Methylsuccinic acid**	n.d	n.d
16. **Phenylacetylglutamine**	n.d	n.d
17. **Pyroglutamic acid**	X	X
18. **Suberic acid**	n.d	n.d
19. **Uracil**	X	X
20. **2-hydroxyglutaric acid**	X	X
21. **3-hydroxy-3-methylglutaric acid**	X	X
22. **3-hydroxyisobutyric acid**	X	n.d
23. **Lactic acid**	X	X
24. **Succinic acid**	X	X
25. **3-hydroxyvaleric acid**	X	n.d

(X): metabolite detected. (n.d): not detected.

Table 4.2 summarizes the organic acids detected in urine samples in previous studies as well as organic acids found uniquely in the respective samples when using two different sample preparation methods. Eight of the metabolites were not detected in either of the samples preparation assays, 13 were detected in both assays, and four of the metabolites were only detected in the deproteinised urine assay.

A number of reasons can be given as to why the deproteinised urine assay presented with four additional metabolites compared to the organic acid extraction assay. Firstly, the four metabolites not detected in the organic acid extraction assay may have been a result of insufficient extraction. In the organic acid extraction process two immiscible liquids are used, the organic phase (diethyl ether and diethyl acetate) and the aqueous phase (water). At low pH organic acids will dissolve in the organic phase, while other polar metabolites will dissolve in the water phase. Ideally, the total concentration of a metabolite present in the sample should dissolve in the organic phase, but this may not always be the case. It is however possible that a portion of the metabolite is in the water phase or is lost, when the upper organic phase is removed. This results in a lower concentration of metabolites, just below the detection limit. Secondly, the organic extraction assay also has a lot of sample handling steps. The more steps that are involved in sample manipulation, the more sample loss that can occur. Not only is the deproteinised urine assay less labour intensive, but it also provides a wider coverage of the metabolome.From the results obtained, it was evident that the acetonitrile deproteinisation of urine would be the better choice in terms of repeatability and detection of metabolites.

4.3.2 Biosignature expansion

4.3.2.1 Overview of data quality

Since batch effects and time drifts are always a cause for concern in metabolomics, the quality of the data was evaluated. No batch effects could be detected following data inspection. A heat map was used to visually inspect the data for any outliers.

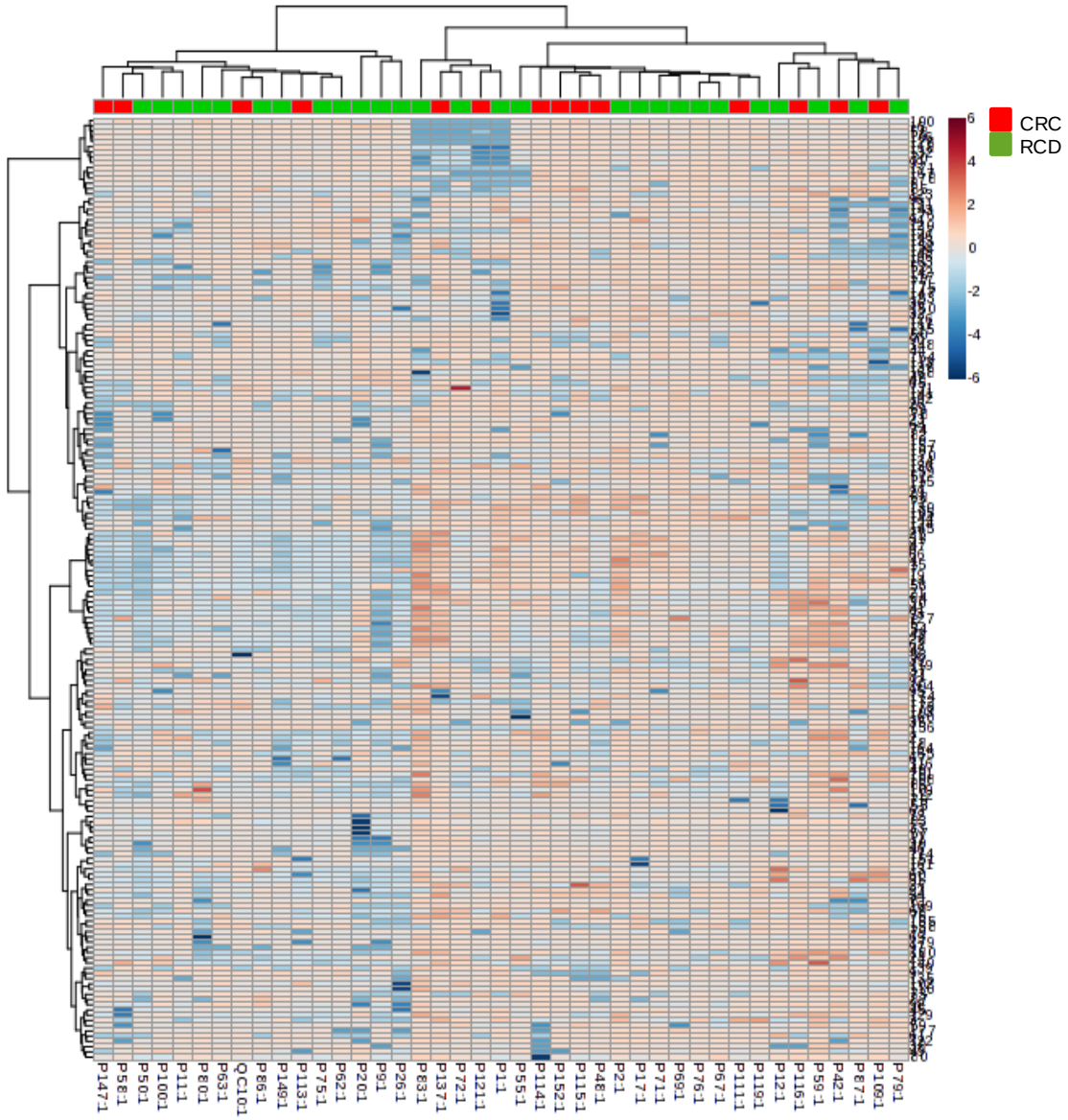

Figure 4.5: Hierarchically clustered heat map of the urinary metabolites of RCD and CRC patients following GC-MS analysis. Inspection of the heat map indicated no visibly clear outliers that had to be removed.

Figure 4.5 shows a heat map generated from the data using hierarchical clustering to sort the samples and features. Each column represents a sample and each row represents a metabolite feature. Heat maps are often used to visualize clinical data across multiple samples to identify patterns in the data (Moon *et al.,* 2009). However, in this study, a heat map was used for visualization of the data to identify possible outliers which could then be removed. Inspection of the heat map indicated no visibly clear outliers.

4.3.2.2 Overview of data prior to feature selection

The original data matrix (prior to data pre-processing) contained a total of 615 features. Following zero filtering, 271 features remained in the dataset. Finally, 180 features remained in the data set following the CV filter, which was the last step in data pre-processing.

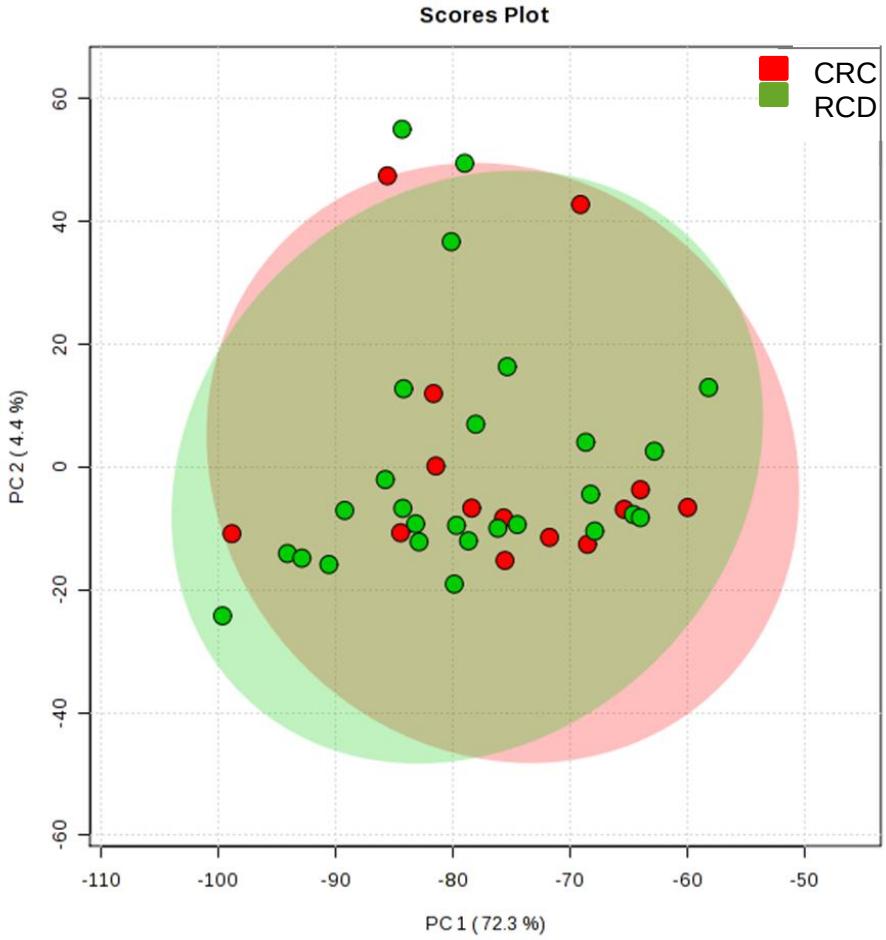

Figure 4.6: PCA score plot of GC-MS data prior to feature selection. Two dimensional PCA score plot of the RCD group (green) and CRC (red) group.

The PCA (Figure 4.6) was constructed by including the 180 features that remained in the data matrix after data pre-processing. In the figure, the RCD group and the CRC group are scattered across the plot. It was expected that the RCD group would not cluster together tightly due to the diversity found within the group and the heterogeneity of RCDs. The CRC group were scattered among the RCD patients, suggesting that these groups had similar metabolite profiles.

Taking into consideration that the GC-MS allows the detection of mainly small compounds belonging to the primary metabolism, it can be assumed that the primary metabolism of these patients is similar (Gullberg *et al*., 2004). This is not unusual since the CRC patients were regarded as potential RCD patients during clinical evaluation, based on their phenotype. Since metabolomics is regarded as the omics field that is closest to the phenotype, one would expect minor differences between these groups on metabolite level (Dettmer *et al*., 2006; Dunn, 2008). With very similar overall metabolic profiles between these groups, it was necessary to perform feature selection in order to identify discriminating metabolites or metabolite features, and to unmask the relevant biological variance.

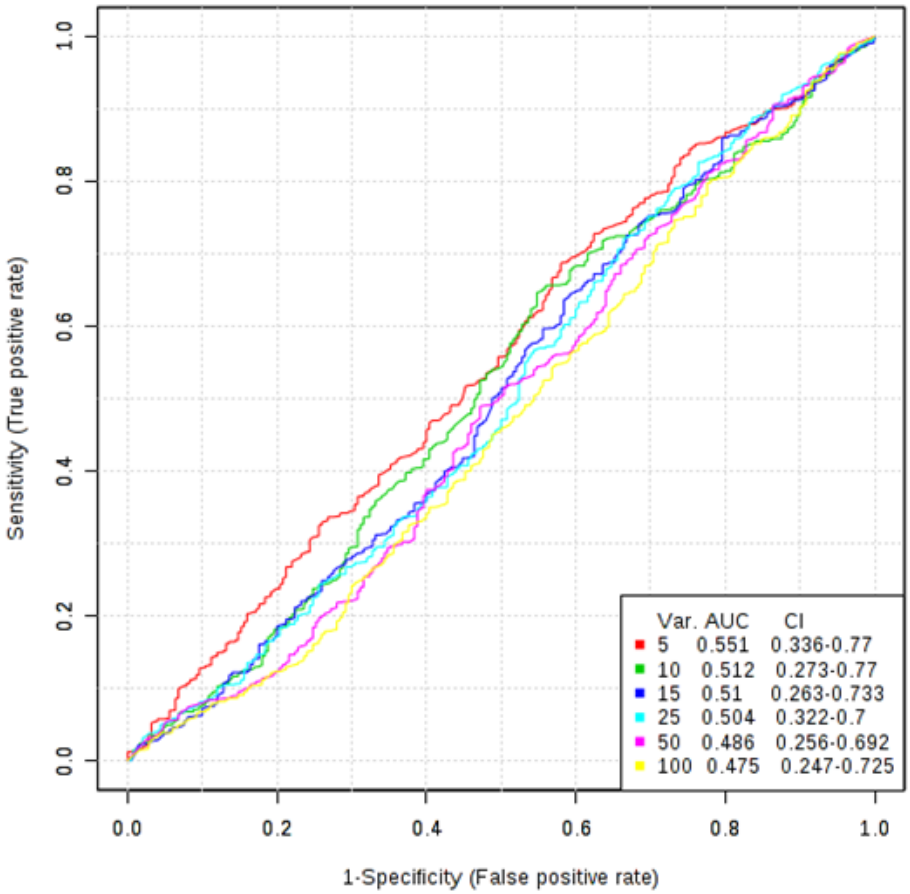

Figure 4.7: Multivariate AUC of GC-MS analysis prior to feature selection. In this figure the confidence and performance intervals of a different number of selected variables are given. Var: variables, CI: confidence interval and AUC: area under the curve**.**

Figure 4.7 illustrates the ROC curves generated for different subsets of metabolites. These curves can be described as a two-dimensional illustration of classifier performance. As seen in Figure 4.7, the six AUC are displayed representing 5, 10, 15, 25, 50 and 100 features.

The AU-ROC will always have a value of between 0 and 1, with the latter being the best possible classifier. The results indicated that the greater the number of features or variables, the smaller the area under the curve (lower overall accuracy of the test). This was in contrast to the LC-MS data where the model became more discriminative, the more variables that were used (Venter *et al*., 2015). This may indicate that most of the relevant variance lies within the secondary metabolism (mostly detected with LC-MS). With the GC-MS data, the AUC became larger when fewer features were used, albeit still very low.

A major challenge in biomarker discovery is how to extract relevant information from the data that is obtained. Due to the highly multivariate nature of metabolomics data, chemometrics is required to find significant information in the data, such as relevant metabolites (Smolinska *et al*., 2012). Feature selection is crucial for determining the list of features that provide the best classification performance on a predefined group of samples (Christin *et al*., 2015). From a clinical point of view, an ideal biosignature would be one that could discriminate between two groups, with the least amount of variables. Feature selection that utilizes various multivariate and univariate statistical tests, identifies the best candidates for a biosignature that can differentiate between two groups.

4.3.2.3 Multivariate and univariate feature selection

Venn diagrams were used as an illustration of the number of features that were detected with each test, and to depict the relationship between these features that were found using different statistical tests. The Venn diagram was not used as a feature selection tool, a common practice in metabolomics, where the features found in the centre of the diagrams are regarded as the most important features. Figure 4.8A illustrates the Venn diagram that contained the features found when using univariate statistical methods. Figure 4.8B illustrates the Venn diagram that provided a summary of features found when using multivariate statistical analyses.

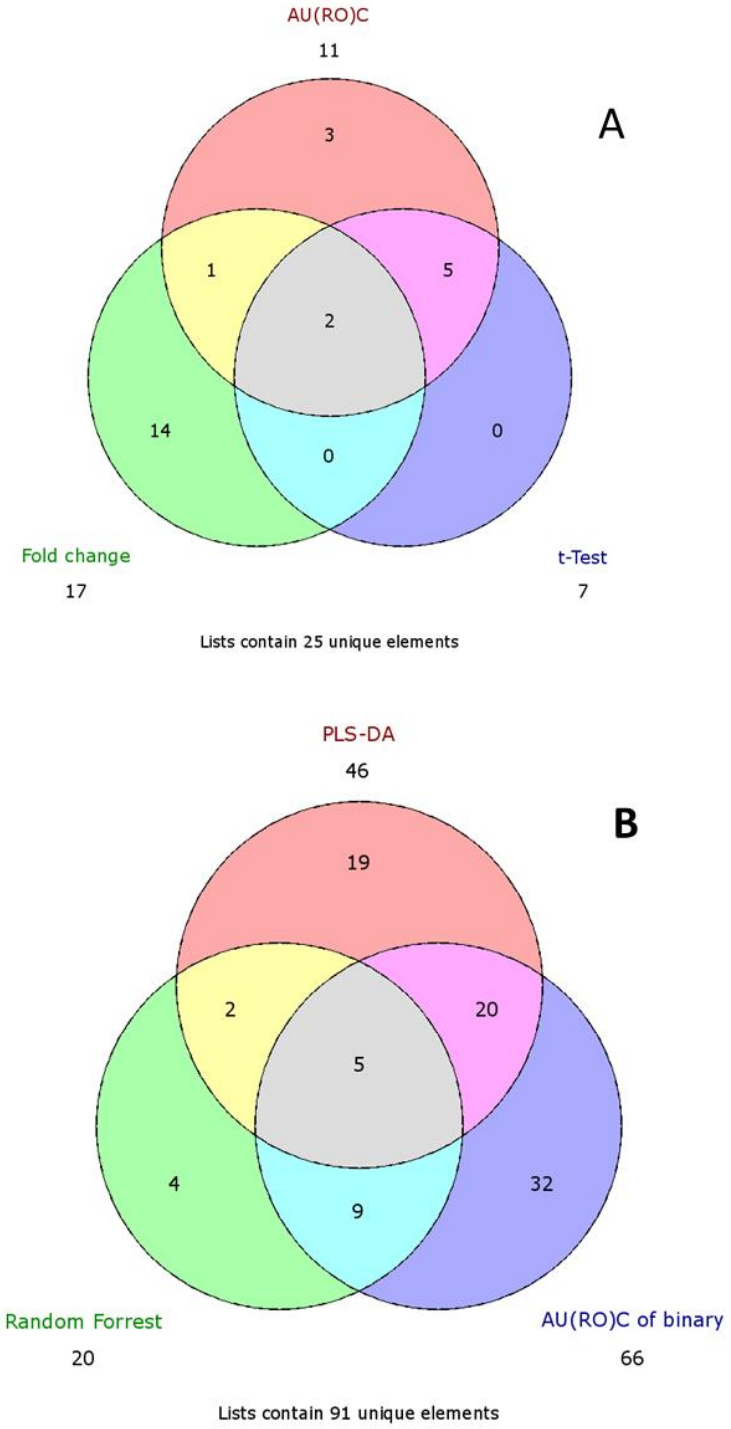

Figure 4.8: Venn diagram illustrating the significant features found with multivariate and univariate statistical methods. A) Represents features found using univariate statistical analyses. B) Represents the features found when multivariate statistical methods were used.

The features of significance found with the Student's *t*-test, fold change and area under the ROC curve (univariate methods) were used to construct the Venn diagram of Figure 4.8A. A total of 25 features, that could be considered as important, were detected using univariate

methods. For the area under the ROC curve, a total of 11 features were found, while 17 features were found when using fold change. Seven features were detected using the Student's *t*-test. The features of significance found in the PLS-DA, bivariate area under the ROC curve and random forest (multivariate methods) were used to construct the Venn diagram of Figure 4.8B. A total of 91 features, which could be regarded as important, were found using multivariate methods. When using PLS-DA, a total of 46 features were found, 66 features using the bivariate area under the ROC curve and 20 features when random forest was used. Since many statistical tests have flaws and some bias, it was decided to use several tests to exploit the different advantages and powers of each test, in the process of feature selection (Saccenti *et al*., 2014).

4.3.2.4 Feature ranking

After multiple statistical tests, the average rank of each feature across all the tests was determined, with a low rank being significant and a high ranking value indicating no importance. With this approach, a feature that was important in most of the tests had a low average rank. By using the average ranking of the features, the graph in Figure 4.9 was constructed.

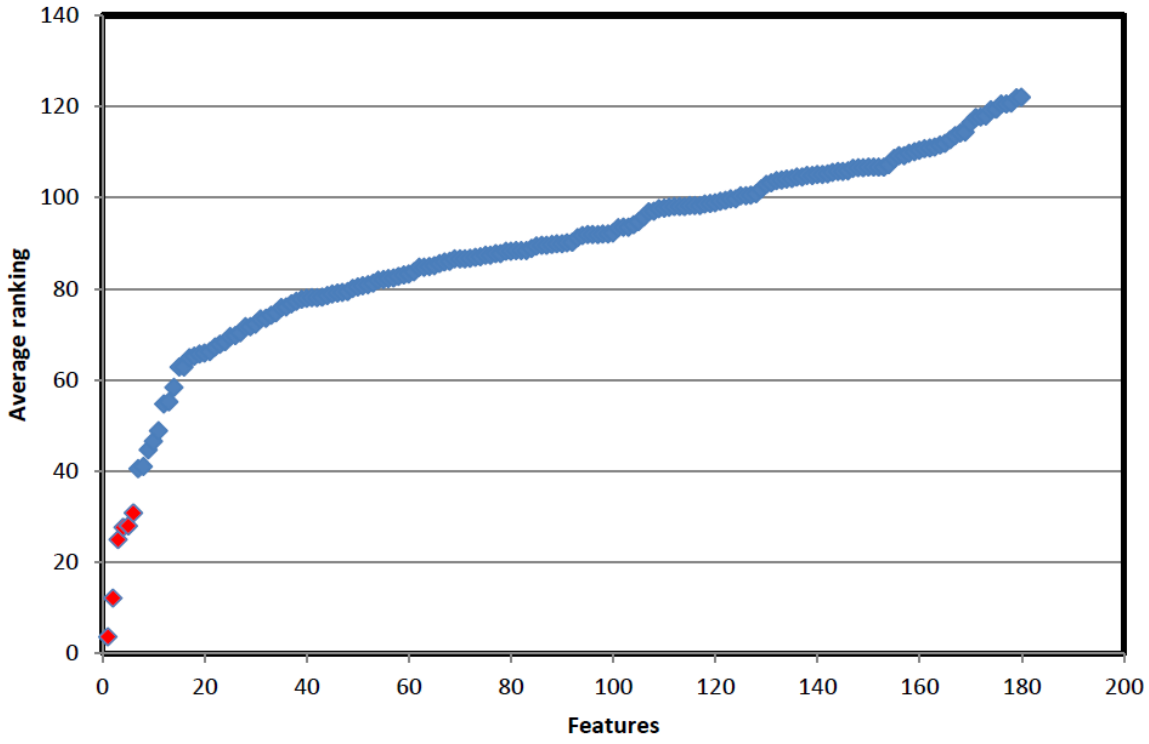

Figure 4.9: Average ranking of important GC-MS features. On the basis of their average ranking, the top six features (in red) presented with the best values.

In Figure 4.9, the average ranking position is plotted against the feature. The graph was used as a guideline to visually select the best possible features that could form part of the improved biosignature. There is no guideline with regards to the number of features that should be included in a biosignature. Considering the results from the multivariate ROC (Figure 4.7) prior to feature selection, the least amount of features gave the best discriminative model. In the above graph it is clear that the six features are ranked at the top of the list and plotted some distance away from the next group of top ranked features. The six top ranked features are listed in Table 4.3.

Table 4.3: GC-MS features for the improvement of the biosignature. The top six features selected for the improved biosignature. The assigned name of the feature and the results of the selected statistical test values (*t*-test *p*-value, PLS-DA VIP value and AUC) are included in this table.

Rank	Name	*t*-test (*p*-value)	Fold change	PLS-DA (VIP)	AUC
1	Oxalic acid	0.001	0.469	2.814	0.725
2	Unknown 1	0.045	2.304	4.569	0.683
3	Ribitol	0.035	-	2.260	0.683
4	Unknown 2	0.032	-	3.596	0.669
5	Citric acid	0.015	-	0.989	0.759
6	β-Pseudouridine	0.021	-	1.054	0.709

4.3.3 Discriminative power of top ranked features

The discriminative power of the six top ranked features was evaluated using a PCA score plot. This gave an indication of the ability of these features to distinguish between the RCD patient group and the CRC group.

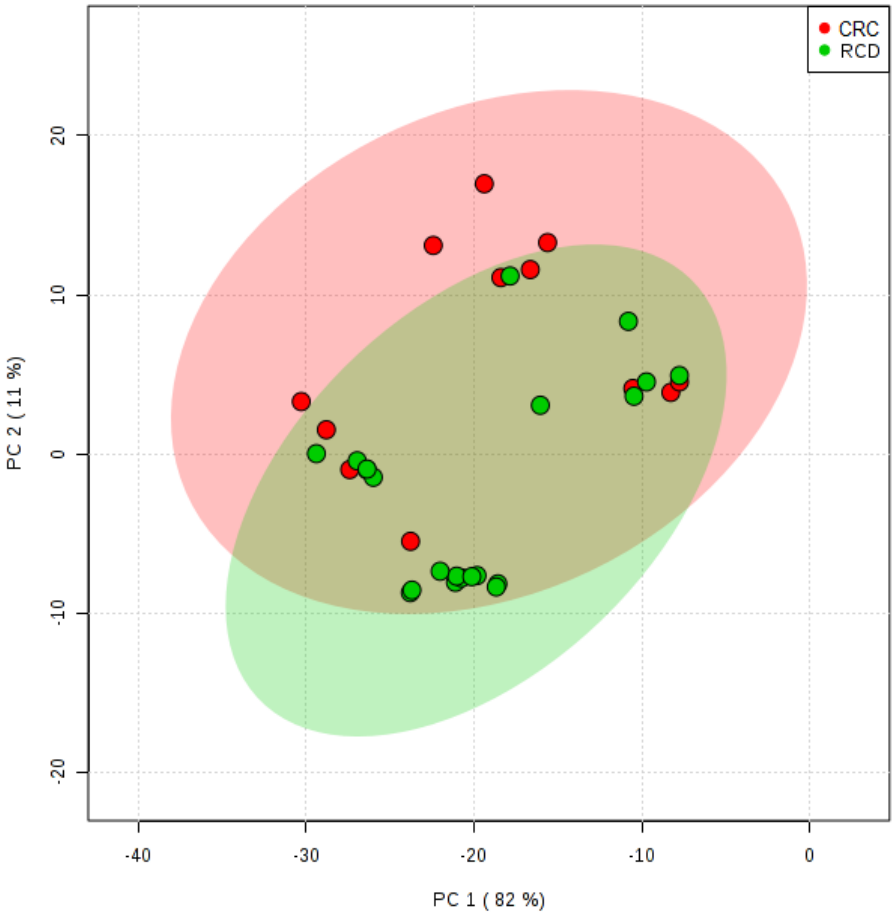

Figure 4.10: PCA score plot of the six top ranked features found with GC-MS analysis. Separation of the RCD patient group (green) and the CRC (red) group, with the use of the six top ranked features, is displayed.

Figure 4.10 shows the PCA score plot of the RCD patient group and the CRC group, when using only the six top ranked features. In Figure 4.10 the RCD group and CRC group are scattered across the plot, with no clustering of groups and no separation between these two groups. As mentioned previously, GC-MS analyses allow the detection of mainly small compounds belonging to the primary metabolism, so it can be assumed that the primary metabolism of these patients is similar. Even after feature selection, the PCA score plot indicates that the six top ranked features were not able to discriminate between these two groups.

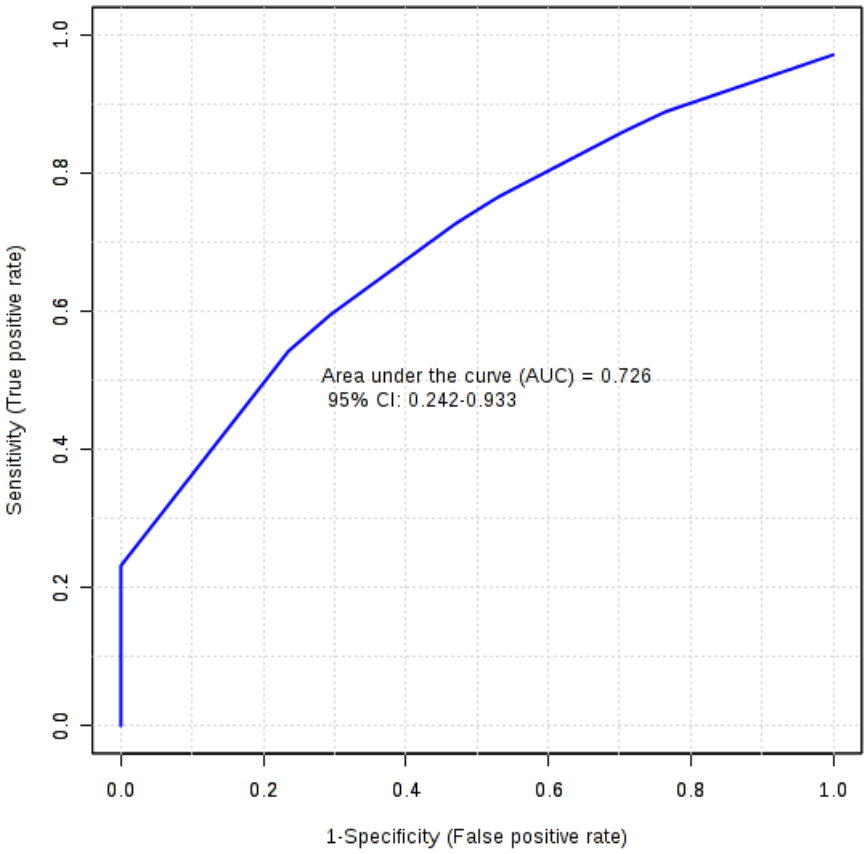

Figure 4.11: ROC curve for the top six ranked features found with GC-MS analyses. The curve displays the results of the evaluation of the top six ranked features with regards to sensitivity and specificity.

Figure 4.11 displays the discrimination power of the six top ranked features found with GC-MS analyses. The AUC of 0.726 (with a 95% confidence interval that ranged between 0.242 and 0.933), indicated that this set of features tested poorly on specificity and sensitivity.

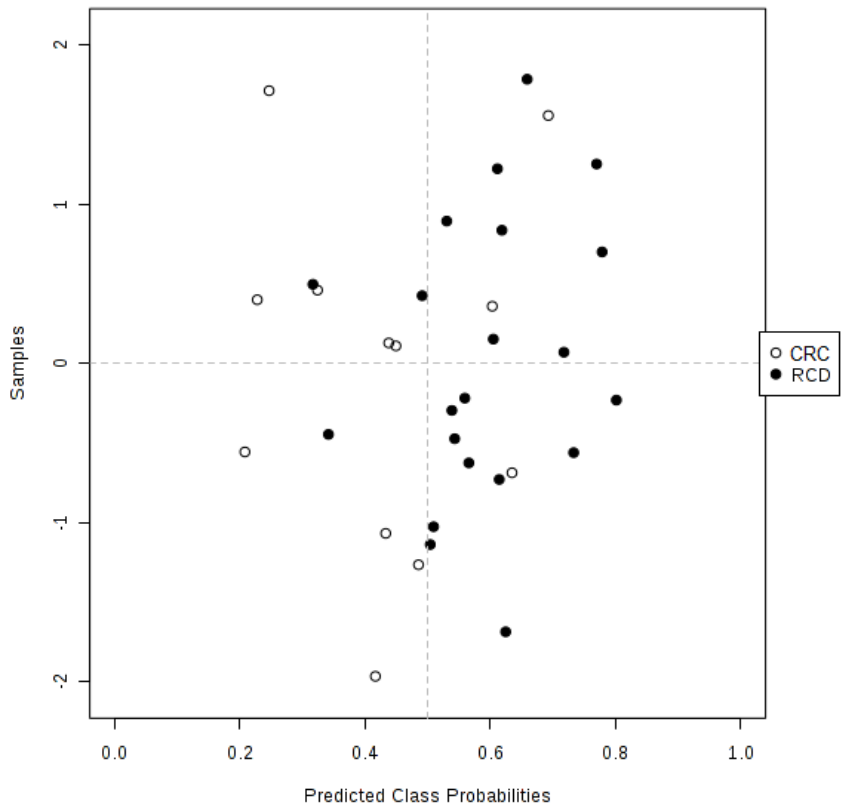

Figure 4.12: Average predicted class probability of the RCD group and CRC group over 100 cross validations for GC-MS. The predicated class probabilities for each sample using the six top ranked GC-MS features are shown. The black dots (●) represent the RCD patient group, while the white dots (○) represent the CRC group.

Figure 4.12 shows the average predicted class probability of each sample, across 100 validations. A balanced sub-sampling approach was used by the validation algorithm with a classification boundary in the centre (0.5). The probability score was calculated and ranged from 0 to 1. When a signature is able to classify the samples 100% accurately in the 100 cross validation sub-samples, then the samples of the one class are distributed together on one side of the 0.5 classification boundary, and the other group on the other side of the boundary. In this figure, the samples were distributed on both sides of the classification boundary, indicating that the combined six top ranked features found with GC-MS analysis, were unable to classify samples accurately in the 100 cross validation sub-samples of the original data.

4.4 Summary

As part of the expansion part of this study, a standardized GC-MS method was developed. For repeatable and reliable analysis of biological samples, method standardization was performed. Minimum sample preparation was compared with organic acid extraction in order to find a suitable sample extraction method. Both methods were evaluated based on method variance and metabolome coverage, with minimal sample preparation performing the best in both cases.

The standardized method was implemented to analyse biological samples in order to identify additional features that could form part of the proposed biosignature of Venter *et al.* (2015). Following statistical analyses, six features were identified when analysing urine samples on the GC-MS. The discrimination power of the six top ranked features was evaluated using a PCA score plot and ROC curves. Results indicated that these sets of features were unable to discriminate between the two groups.

It should be noted that the metabolite markers in a biosignature perform as a group, and although these six top ranked features of GC-MS analysis were not able to distinguish between the two groups, it does not mean that they are not of value. These features together with the verified features were considered for the improved biosignature (Chapter 6). As mentioned previously, MS and NMR are techniques most commonly used in metabolomics. For this reason it was decided to include NMR analyses on this sample group in an attempt to identify features that could be added to the improved biosignature. This will be discussed in the following chapter.

CHAPTER 5

NUCLEAR MAGNETIC RESONANCE

5

5.1 Introduction

NMR, GC-MS and LC-MS are powerful analytical platforms used in metabolomics to generate metabolomic profiles. All of these methods have advantages and disadvantages. For instance LC-MS and GC-MS are more sensitive than NMR and may detect metabolites that have a concentration below the detection limit of NMR. NMR, however, is untargeted, non-destructive and requires limited sample preparation. Together, these three techniques are used to maximize metabolome coverage (Enwas *et al.,* 2013). The sample set used in the study conducted by Venter *et al*. (2015) was not previously analysed using NMR. The decision was made to include NMR analyses on this sample set, in an effort to expand the existing biosignature.

This chapter includes all aspects of the NMR methodology that was used to expand the biosignature. As a starting point, the established method was evaluated for repeatability. Next, the method was implemented on the sample set consisting of RCD patients and CRC. Lastly, the statistical methods and results will be discussed.

5.2 Materials and methods

5.2.1 Reagents

Potassium dihydrogen phosphate (KH_2PO_4) (Cat # 229806), Potassium hydroxide (KOH) (Cat # P1767) and sodium azide (NaN_3) (Cat # S2002) were purchased from Sigma-Aldrich. Trimethylsilyl-2,2,3,3-tetradeuteropropionic acid (TSP) (Cat # 269913) and deuterium oxide (D_2O) (Cat # 435767) were purchased from Merck chemicals.

For the preparation of 100 mL of 1.5 M KH_2PO_4 buffer solution, 20.4 g of the salt was dissolved in 80 mL of D_2O. TSP and NaN_3 were weighed off separately, at 100 mg and 13 mg respectively, and dissolved in 10 mL of D_2O. The combined solution was mixed using sonication. The pH was adjusted to 7.4 by adding KOH and the solution was then transferred to a 100 ml volumetric flask and topped up with D_2O (Dona *et al*. 2014).

5.2.2 Instrumentation

Spectra were obtained using a Bruker Avance III HD NMR spectrometer equipped with a triple-resonance inverse (TXI) 1H {15N, 13C} probe head and x, y, z gradient coils. Samples were measured at 500 MHz. 1H spectra were acquired as 128 transients in 32K data points, with a spectral width of 6002 Hz. Water resonance was pre-saturated by single-frequency irradiation during a relaxation delay of 4 seconds, with a 90° excitation pulse of 8 µs. Shimming was performed automatically on deuterium signal. Resonance widths for TSP and metabolites were less than 1 Hz. Fourier transformation as well as phase and baseline correction was done automatically. Software used for these purposes was Bruker Topspin and Bruker AMIX (Ellinger *et al*. 2013).

5.2.3 Sample preparation

As mentioned previously, NMR requires limited sample preparation. Urine samples (aliquots of 500 µL per sample) were thawed overnight (4°C). The samples were centrifuged at 25 000 x *g* for 5 minutes at 25°C, followed by transfer of the supernatant (450 µL) into a clean microcentrifuge tube. The buffer solution (45 µL) was added and the mixture vortexed. The sample was then transferred to the 5 mm NMR tubes (Ellinger *et al*. 2013).

5.2.4 Data processing

Data pre-processing can be described as the intermediate step that is used to transform raw data into a data set fit for data analysis. The resulting datasets are more comparable and improve data analysis. Pre-processing steps for NMR spectra usually consist of baseline correction, alignment, binning and normalization.

5.2.4.1 Baseline correction

Baseline removal is usually the first step in data pre-processing, in order to avoid baseline distortions that affect quantification and statistical analyses (Smolinska *et al*., 2012). Baseline correction was done automatically with Bruker AMIX.

5.2.4.2 Alignment

Peak shifts that exist between spectra may cause variations that obscure the discovery of important patterns in spectra. There are a number of factors that can result in shifts, for instance instrumental factors, changes in pH, changes of salt concentration and a common cause is that of an overall dilution (Smolinska *et al*., 2012).

5.2.4.3 Binning

A NMR spectrum contains approximately 22 000 data points (variables) after excluding certain peaks, such as urea, internal references and water. Binning is a data pre-processing method used for NMR data processing to reduce data dimensionality. The spectra are divided into segments, also referred to as bins, and for each bin the total area is calculated to represent the original spectra. Spectral binning is usually in the region of 0.01 - 0.04 ppm, and in this study the most commonly used bin size as described by Craig *et al.* 2006 of 0.04 ppm was selected. Only bins that were seen as important, using statistical analyses were identified and re-integrated where necessary. Each of these bins can be referred to as features.

5.2.4.4 Normalization

The normalization of data can be seen as a critical step in any statistical analysis, since normalization allows objective comparisons between samples. The main aim of this step is to make samples in a data set more comparable (Emwas *et al*., 2013). All spectra were normalized using the creatinine value of each sample.

5.2.5 Method evaluation

Even though NMR is a highly repeatable and reproducible analytical technique, a repeatability study was still conducted to ensure accurate and reliable data acquisition when analysing patient samples (Fages *et al*., 2013). The repeatability of the analytical technique was not the concern but rather the analysis as a whole, which included sample preparation. Once again a CV curve was used to evaluate the repeatability of the measurements.

5.2.5.1 Pooled urine sample analysis

A pooled urine sample, representing RCD patients in the sample set, was used (Section 3.2.4). The pooled urine sample, with a creatinine value of 4.359 mmol/L, was obtained and stored in a -80°C freezer in aliquots of 500 µL. Five of these pooled urine samples were thawed and sample preparation was performed as described in Section 5.4.

5.2.6 Biosignature expansion

5.2.6.1 **Biological samples**

Ethical approval, patient selection and experimental groups were described in Section 3.2.3 Urine samples were analysed according to the procedure described in Section 5.2.3. The samples were analysed in two batches. Quality control samples were included to evaluate and compare the quality and performance of a particular analytical process. As mentioned previously, QCs are included to determine and correct any inconsistencies that can occur during sample preparation or analysis. Due to the high reproducibility of NMR analysis, only two QC sample were added to each batch, one at the beginning and one at the end. The layout of the run order is shown below.

The run order was as follows:

Batch 1: QC1, Samples 1-20, QC2

Batch 2: QC3 Samples 21-41, QC4

5.2.6.2 Statistical analysis

An identical statistical approach to that described in Section 4.2.6.2 was used to find features with high discrimination power. The ranked features/bins were identified and re-integrated.

5.3 Results and discussion

5.3.1 Method evaluation

In this case, CV distributions were used to compare the variability of five replicates of a pooled urine sample.

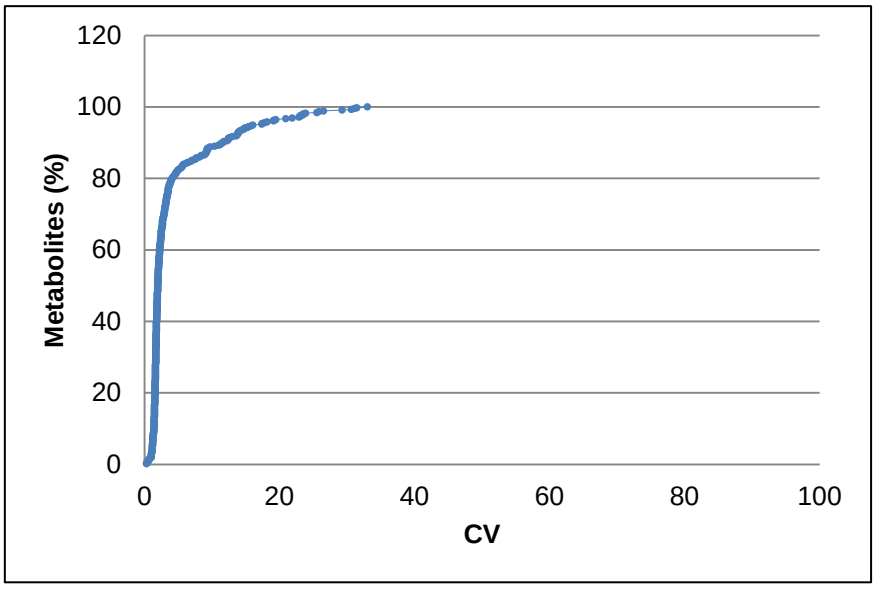

Figure 5.1: CV distribution of NMR analysis. The y-axis shows the percentage of variables (or NMR bins) over CV (x-axis).

Figure 5.1 represents the CV distribution of the features (bins) found in the five replicas of the pooled urine sample. The graph shows that about 89.3 % of the features have a CV distribution of <10 %. %. Table 5.1 provides a summary of the CV distribution values of each extraction method.

Table 5.1: Summary of the CV distribution values

CV range	Percentage of variables in range
<10	89.3 %
10-20%	7.4 %
20-30%	3.3 %
30-40%	0.7 %
>40-50%	0%

By looking at the CV distribution, it was clear that the repeatability of this analysis was good, with almost 90% of features having a CV value within the range of 0-10%. In doing this repeatability study, the total process (from sample preparation to analysis) was evaluated. From the results obtained it was concluded that the entire procedure, including sample preparation, was repeatable and that the analysis of patient samples would yield good quality data.

5.3.2 Biosignature expansion

5.3.2.1 Overview of data quality

Since the sample set was divided into two batches and was analysed over a two day period, it was necessary to inspect the data for between-batch effects, despite NMR being regarded as a highly reproducible technique (Fages *et al.*, 2013). For this purpose a scatter plot and heat map were used to visually inspect the data.

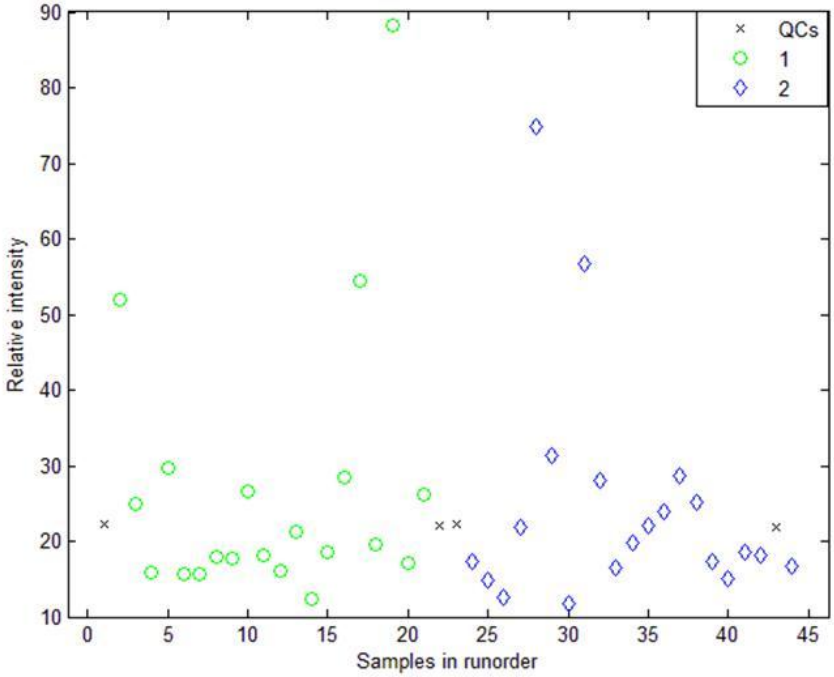

Figure 5.2: Sequential total signal scatter plot of NMR analysis. The two batches that were analysed are visible; with batch one represented by green dots and batch two represented by blue dots. The black X represents the QC samples that were analysed at the beginning and end of each batch.

The total signal of each compound (after normalization with creatinine) was calculated, to construct a scatter plot. The scatter plot, depicting the sum of the features present in each sample, was used to visually evaluate the quality of the data. Additionally, the scatter plot of several random features (bins) was also evaluated to determine whether their respective intensities in the QCs samples were similar (results not shown). The samples were plotted in the order that they were analysed. When assessing data quality, the focus was placed on the QC samples and the overall distribution of the remaining samples. In Figure 5.2, the QC samples were distributed in a straight line, indicating that no obvious variation or visible time-related drift occurred during sample analysis. The samples that appeared at the top of the figure were due to inherit biological variance, and not technical variation. The results in this figure indicated that the quality of the data was good and could be further studied.

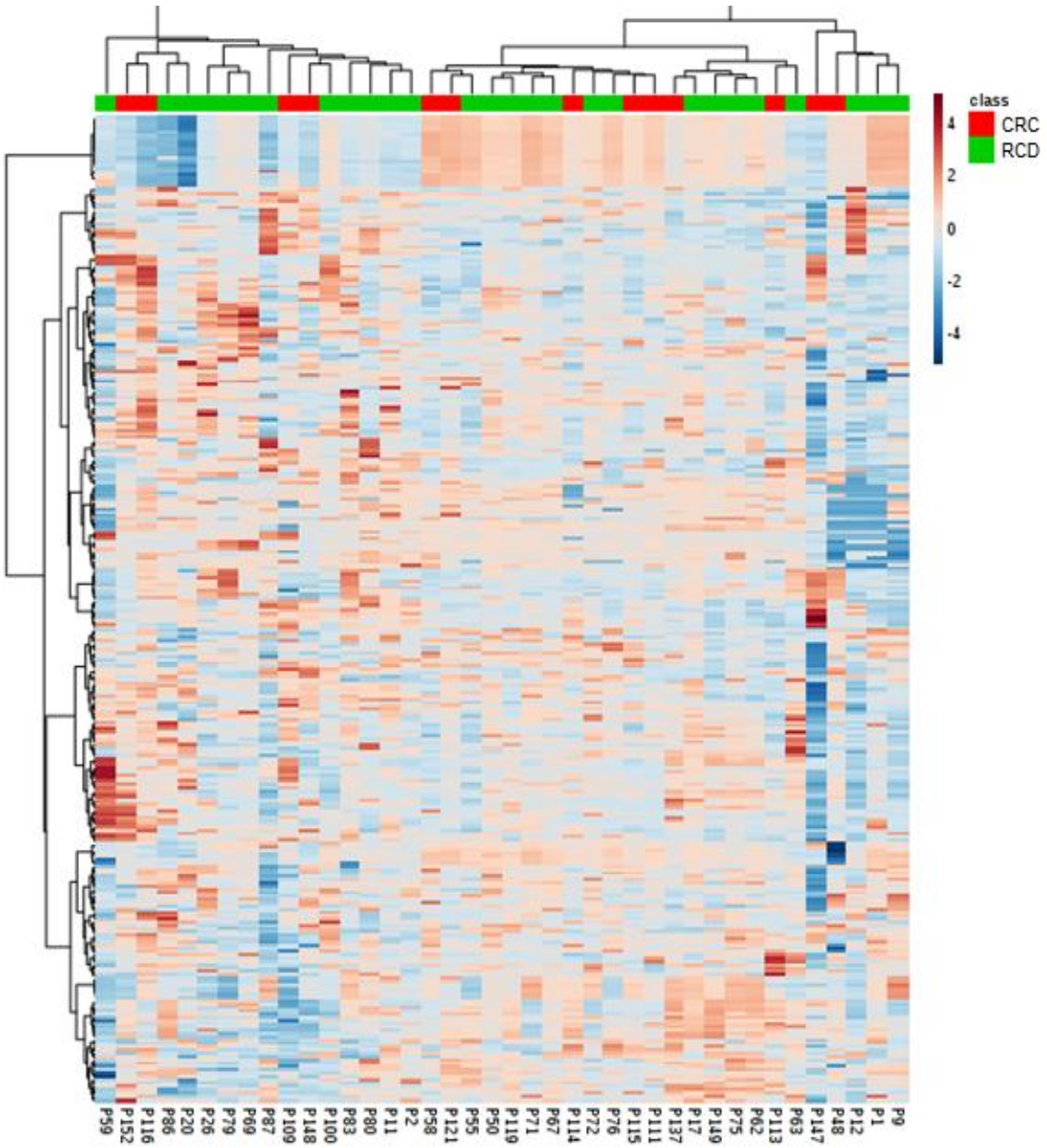

Figure 5.3: Hierarchically clustered heat map of features found with NMR analysis of RCD and CRC patients. Heat maps enable the visualization and removal of outliers.

Following data clean up, the heat map in Figure 5.3 was constructed with hierarchical clustering to sort the samples and features. As previously mentioned, the heat map was used to identify possible outliers. After inspection of the heat map, no clear outliers were identified.

5.3.2.2 Overview of data prior to feature selection

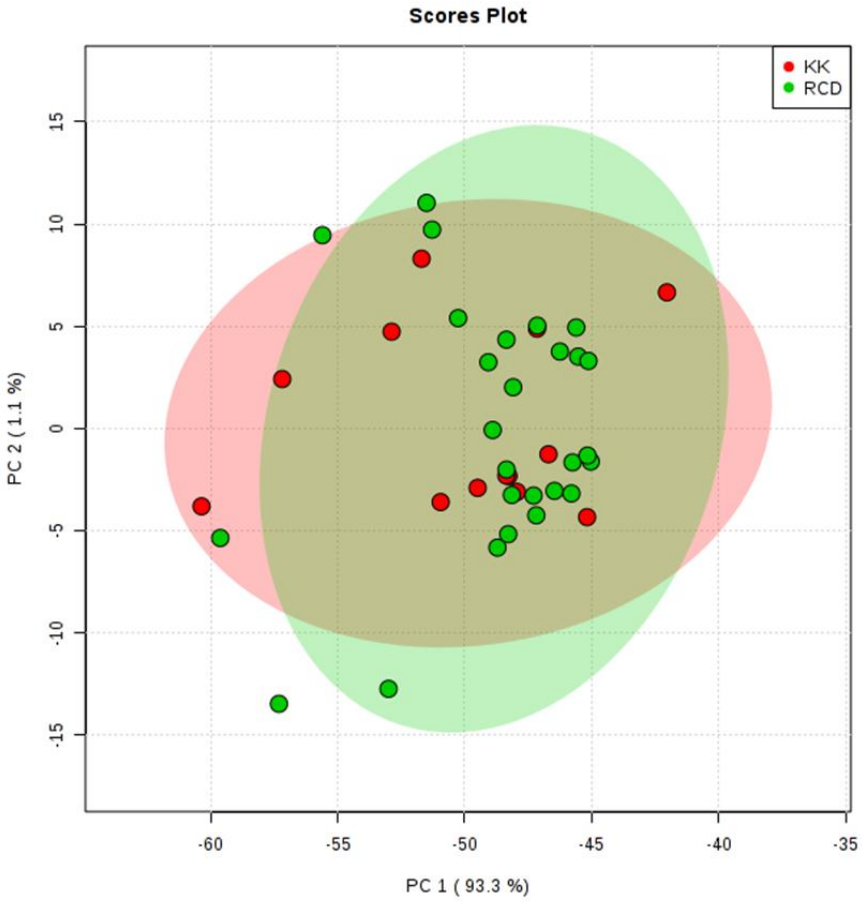

Figure 5.4: PCA score plot prior to feature selection for NMR analysis. Two dimensional PCA score plot of the RCD group (green) and CRC (red) group displayed above.

The PCA score plot above (Figure 5.4) was constructed by including all features detected with NMR analysis. The RCD group and CRC group is once again scattered across the PCA score plot, as seen in the PCA that was constructed using all the features detected with GC-MS analysis.

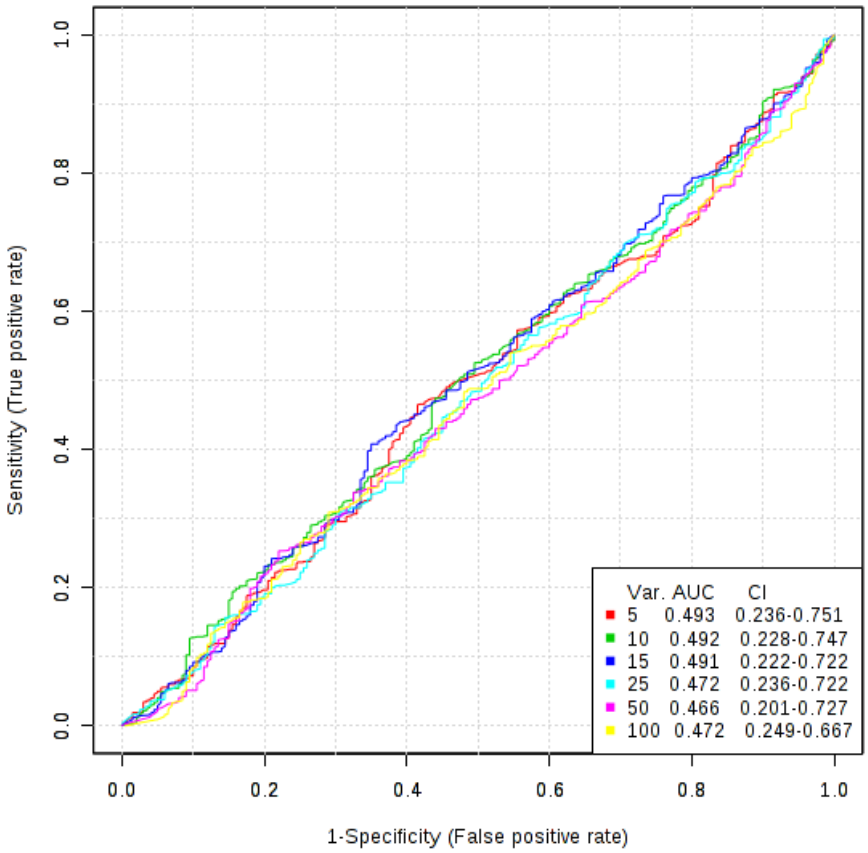

Figure 5.5: Multivariate AUC prior to feature selection for NMR analysis. In this figure the confidence and performance intervals of a different number of selected variables are displayed.

The results of the multivariate ROC that was performed for all metabolites identified using the NMR corresponded to the findings from the GC-MS data, where less features used results in a larger AUC. Taking this into consideration the difference in the amount of features detected between the three platforms, could explain why less feature results in better AUC when fewer metabolites are detected. NMR that depends on spectral resolution has the ability to specifically detect 40-100 metabolites in a urine sample. In contrast, MS methods could detect more than 500 metabolites, with LC-MS capable of detecting more metabolites than GC-MS (Enwas *et al*., 2013). The wider range of metabolites detected with the LC-MS provided more metabolites that could contribute to the variation seen between these two groups and thus result in better classifier performance. However, as mentioned previously, an ideal biosignature would be one that could discriminate between these two groups with the least number of features.

5.3.2.3 Multivariate and univariate feature selection

As mentioned previously, Venn diagrams were used to illustrate detected features and the relationship between features that were detected using different statistical tests. Figure 5.5A illustrates the Venn diagram containing the features found when using univariate statistical methods. Figure 5.5B illustrates the Venn diagram that provides a summary of features found when using multivariate statistical analysis.

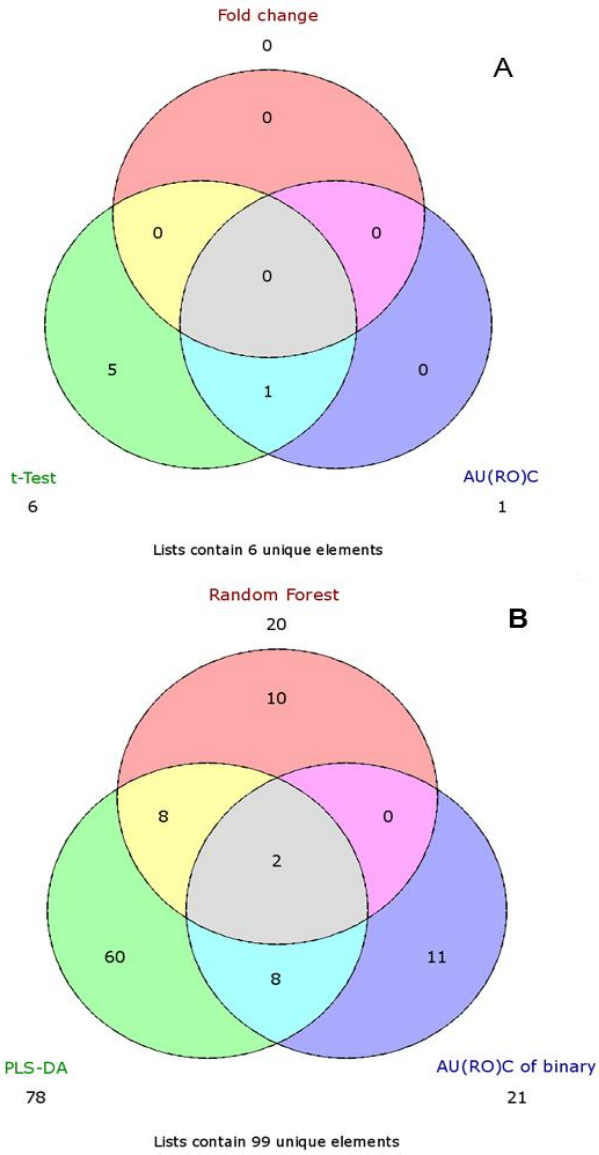

Figure 5.6: Venn diagram illustrating the significant features found in the NMR analysis using multivariate and univariate statistical methods. Figure 5.6A illustrate the Venn diagram that contain the features found when using univariate statistical methods and Figure 5.6B illustrate the Venn diagram that provide a summary of features found when using multivariate statistical analysis.

The features of significance found in the t-test, fold change and area under the ROC curve (univariate methods) were used to construct the Venn diagram of Figure 5.6A. A total of 6 features, which could be considered to be important, were found using univariate methods, where six features were found with the use of a t-Test, one feature using the area under the ROC curve and none using fold change. The features of significance found in the PLS-DA, bivariate area under the ROC curve and random forest tests (multivariate methods) were used to construct the Venn diagram of Figure 5.6B. A total of 99 features, which could be regarded as important, were found using multivariate methods where PLS-DA identified a total of 78 features, bivariate area under the ROC curve found 21 features and 20 features were identified using random forest.

5.3.2.4 Feature ranking & re-integration

Following multiple statistical tests, the average rank of each feature across all tests was determined. The graph in Figure 5.7 was constructed using the average ranking for features. The low rank features (in red) being significant and high rank features (in blue) being of no importance. Therefore, the features that were regarded as important in most tests had a low average rank.

Due to software limitations, it was more practical and allowed for higher-throughput, to extract the NMR data into a matrix of bins rather than compounds (as integration of compound peaks is a manual job). For this reason, the important bins had to be identified and re-integrated before the features could be included in the final candidate list of the biosignature. Four features (bins) performed well in each of the tests and had a low ranking. The four bins were identified and re-intergrated and the information was listed in Table 5.2. NMR peaks were identified using a Bruker spectral library database and an in-house pure compound database.

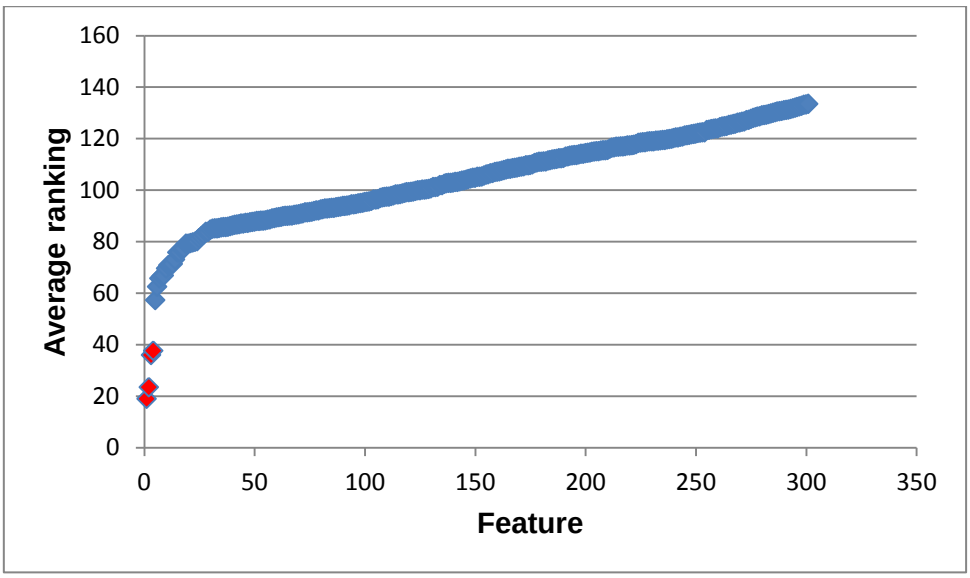

Figure 5.7: Average ranking of important features found with NMR analysis. On the basis of their average ranking, the top four features presented with the best values

Ranking	Name	Feature/bin	*t*-test (*p*-value)	PLS-DA (VIP)	AUC
1	Citric acid	4.6099	0.0116	2.3277	0.7834
2	Methylmalonic acid	3.3099	0.0084	2.737	-
3	Uric acid	2.5699	0.0167	4.3031	-
4	Galactose	3.1099	0.0376	2.0695	-

Table 5.2 shows the results obtained from the various statistical tests performed (before re-integration). Some of the statistical tests, for instance random forest, only provided a ranking where the top 20 features were regarded as important. For this reason, the only statistical test that provided a statistical value was included in the table. In the table there are some missing values for certain features, meaning that the feature performed well across all other statistical tests but not in that particular test. However, statistical analysis revealed that these features had a low rank, thus indicating that the feature could be regarded as significant. It should be noted that the re-integration resulted in fewer differences between the groups. For example, the p-values for two of the four compounds (citric acid and galactose) increased to >0.05. Nevertheless, all four compounds were included in the final list of candidates.

5.3.3 Discriminative power of top ranked compounds

The discriminative power of the four top ranked compounds was evaluated using a PCA. This gave an indication of the ability of these compounds to distinguish between the RCD patient group and the CRC group.

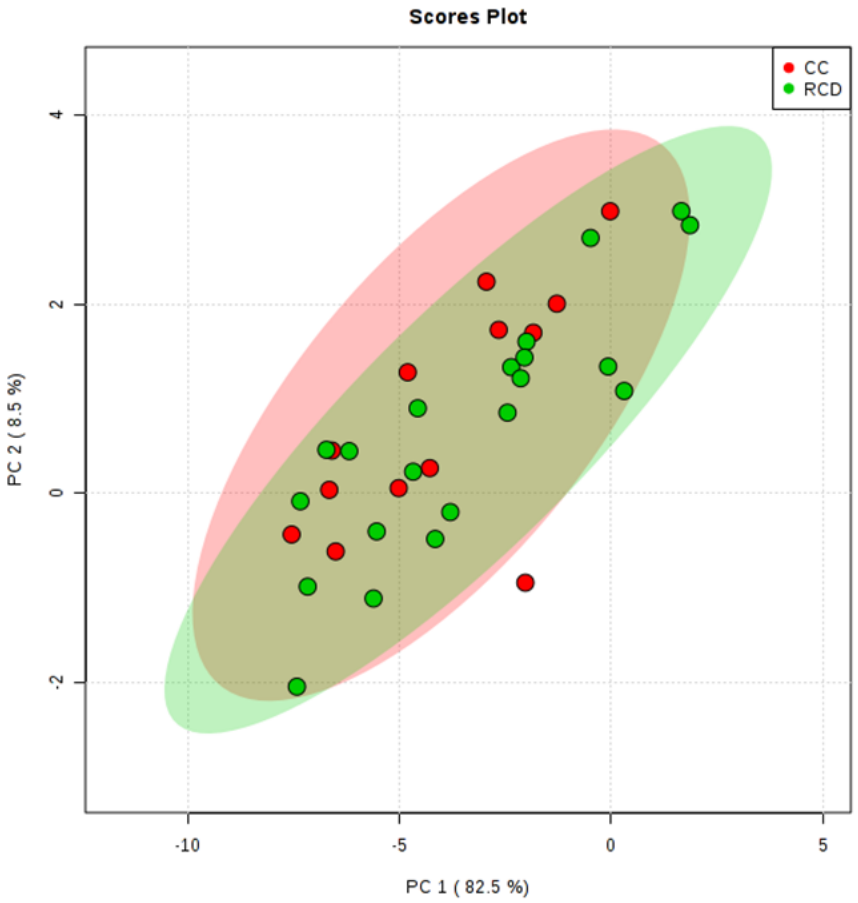

Figure 5.8: PCA score plot of the four top ranked compounds found with NMR analysis. Separation of the RCD patient group and CRC group is displayed, with the use of the six top ranked features.

Figure 5.8 shows the PCA score plot of the RCD patient group and CRC group when using only the four top ranked compounds. In Figure 5.8 the RCD group and CRC group are scattered across the plot, with no clustering of groups and no separation between these two groups. The sensitivity of NMR remains the limitation of this technique, with low concentration features being masked by high concentration features. As mentioned previously, the important features for biological variance are not necessarily the features with the highest concentration, thus resulting in features that could have high discriminative power but which are not detected.

Even after feature selection, the PCA score plot indicated that the combined four top ranked features were not able to distinguish between these two groups. One of the reasons, as mentioned, was that the discriminative power of the top compounds was lost after the important bins were re-integrated. Also, PCA is only a visualization tool which is not often used to classify cases. While the PCA score plot indicated that the cases were indistinguishable from one another, it did not mean that the algorithms used in such cases would give the same result. Algorithms such as logistic regression are commonly used as a classifier, when two or more measurements are combined.

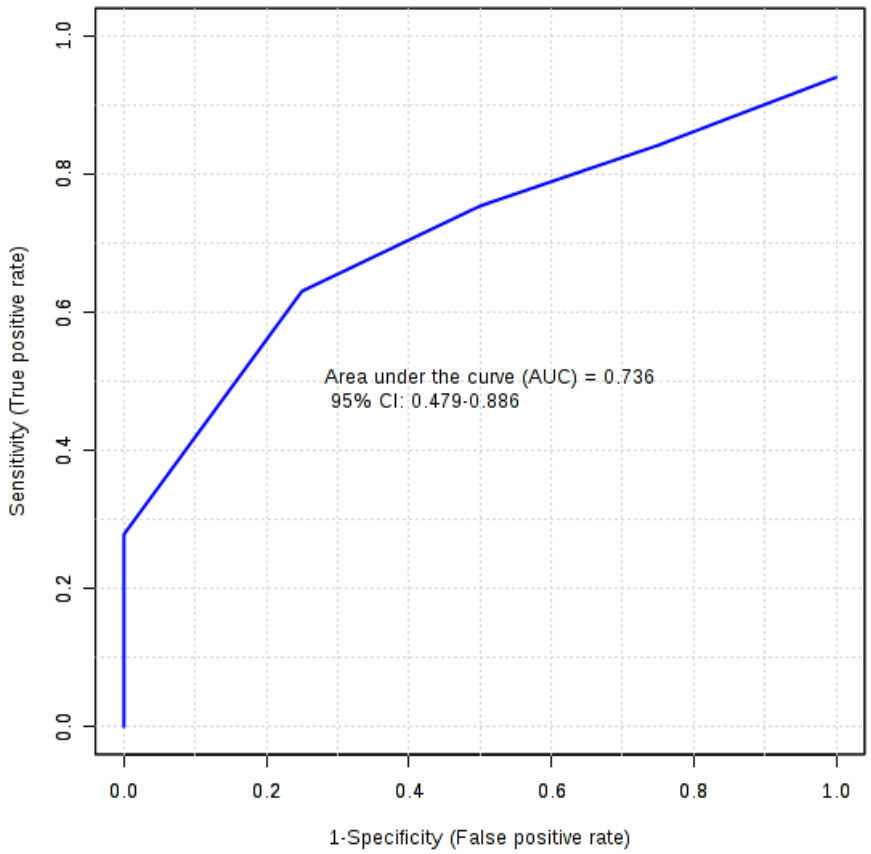

Figure 5.9: ROC curve for the top four ranked compounds found with NMR analyses. The curve displays the results of the evaluation of the top four ranked compounds with regards to sensitivity and specificity.

Figure 5.9 displays the discrimination power of the four top ranked compounds found using NMR analyses. The AUC of 0.736, with a 95% confidence interval ranging between 0.479 and 0.886, indicated that this set of compounds tested relatively poorly on specificity and sensitivity.

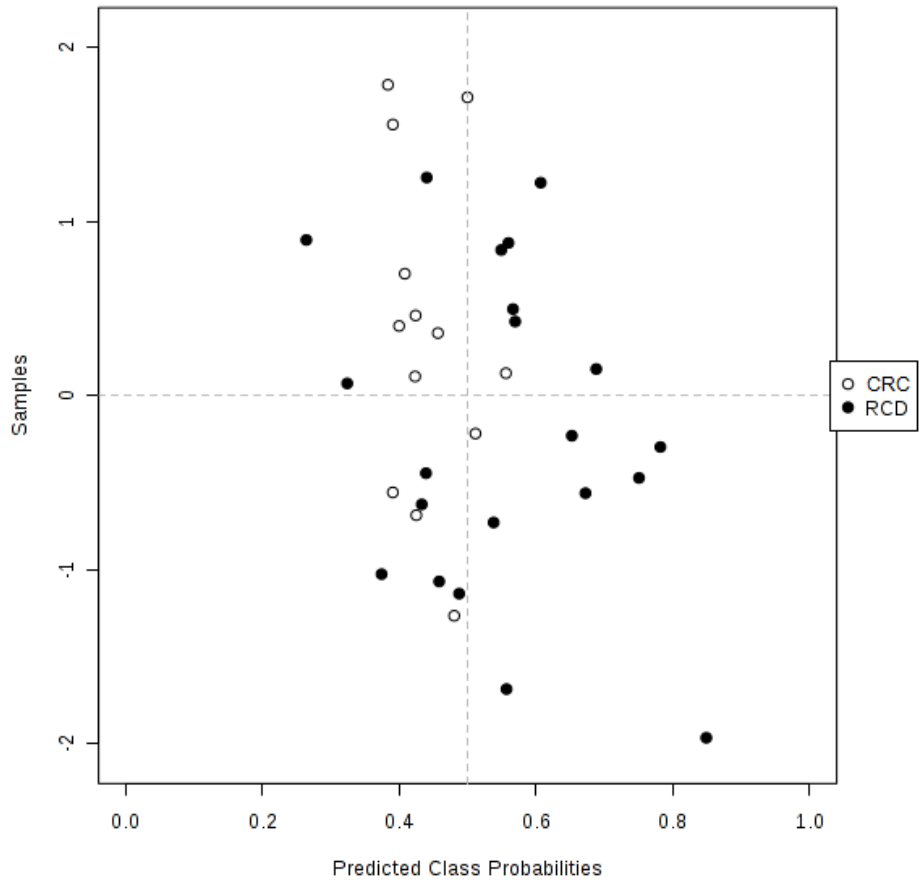

Figure 5.10: Average predicted class probability of the RCD group and CRC group over 100 cross validations. The predicated class probabilities for each sample, using the four top ranked NMR features, are shown. The black dots (●) represent the RCD patient group, while the white dots (○) represent the CRC group

Figure 5.10 shows the average predicted class probability (X-axis) of each sample (Y-axis) across 100 validations. A balanced sub-sampling approach was used by the validation algorithm, with a classification boundary in the centre (0.5). The probability score was calculated and ranged from 0 to 1. When a signature is able to classify the samples 100% accurately in the 100 cross validation sub-samples, then the samples of the one class are distributed together on one side of the 0.5 classification boundary and the other group on the other side. In this figure, the samples are distributed on both sides of the classification boundary, thus indicating that the combined four top ranked features, identified using NMR analysis, were unable to accurately classify samples in the 100 cross validation sub-samples of the original data.

5.4 Summary

As part of the expansion part of this study, NMR analysis was included using the identical sample cohort to Venter *et al*. (2015). Even though NMR is a highly repeatable and reproducible analytical technique, a repeatability study was still conducted to ensure accurate and reliable data acquisition when analysing patient samples. The repeatability of the analytical technique was not the concern but rather the analysis as a whole, which included sample preparation. The repeatability of the procedure proved to be good and was suitable for analyses of biological samples.

Following statistical analyses, four features were found when NMR analyses were performed on biological samples. The discrimination power of the four top ranked features was evaluated using a PCA score plot and ROC curves. Results indicated that the four top ranked features of NMR analyses were unable to distinguish between these two groups, as was the case with the GC-MS analyses.

The verified features, together with top ranked features of GC-MS and NMR, were considered for the improved biosignature. Using statistical analyses, the best possible combination of features were selected to form part of the improved biosignature, which will be described in Chapter 6.

CHAPTER 6

COMPILATION OF AN IMPROVED BIOSIGNATURE

6

6.1 Introduction

Mass spectrometry and NMR are the main technical approaches in generating metabolomic data. By using these two platforms in combination, the metabolome coverage in biological samples can be extended. An example of a study that utilized these two platforms for a metabolomics investigation was Smuts *et al.* (2012) who used targeted LC-MS, GC-MS and NMR analyses to construct a urinary biosignature that was able to distinguish between healthy controls and RCDs. By using multiple analytical instruments, the LC-MS features of the urinary biosignature proposed by Venter *et al*. (2015) were verified and GC-MS and NMR analyses were performed (using the identical sample cohort) in order to discover discriminative features that could be considered for the improved biosignature.

In this chapter, the discriminative power of the verified and top ranked features from all three platforms was evaluated. The results of the statistical analyses of the selected feature and the discriminative power of the improved biosignature are also shown.

6.2 Discriminative power of verified and top ranked features

Following statistical analysis and feature selection the top ranked features of each platform were identified. In Table 6.1 the list of verified LC-MS features together with the top ranked features of GC-MS and NMR analyses are shown. In Table 6.1 the list of verified LC-MS features together with the top ranked features of GC-MS and NMR analyses are shown.

Table 6.1: Verified and top ranked features of respective platforms.

Annotated name	Platform
1. $C_{23}HNO_8S_4$	LC-QTOF
2. $C_{16}H_{26}O4_S$	LC-QTOF
3. $C_{14}H_{22}O_2$	LC-QTOF
4. Androstenol	LC-QTOF
5. Tetrapeptide	LC-QTOF
6. Oxalic acid	GC-MS
7. Unknown 1	GC-MS
8. Ribitol	GC-MS
9. β-Pseudouridine	GC-MS
10. Unknown 2	GC-MS
11. Citric acid	GC-MS/NMR
12. Methylmalonic acid	NMR
13. Uric acid	NMR
14. Galactose	NMR

In Table 6.1 the annotated names of features are given along with the platform used to discover each features. The table contains the verified (LC-MS) and top ranked (GC-MS and NMR) features of the respective platforms. These features were combined and the discrimination power was evaluated. Figure 6.1 indicate the PCA score plot of the features listed in Table 6.1.

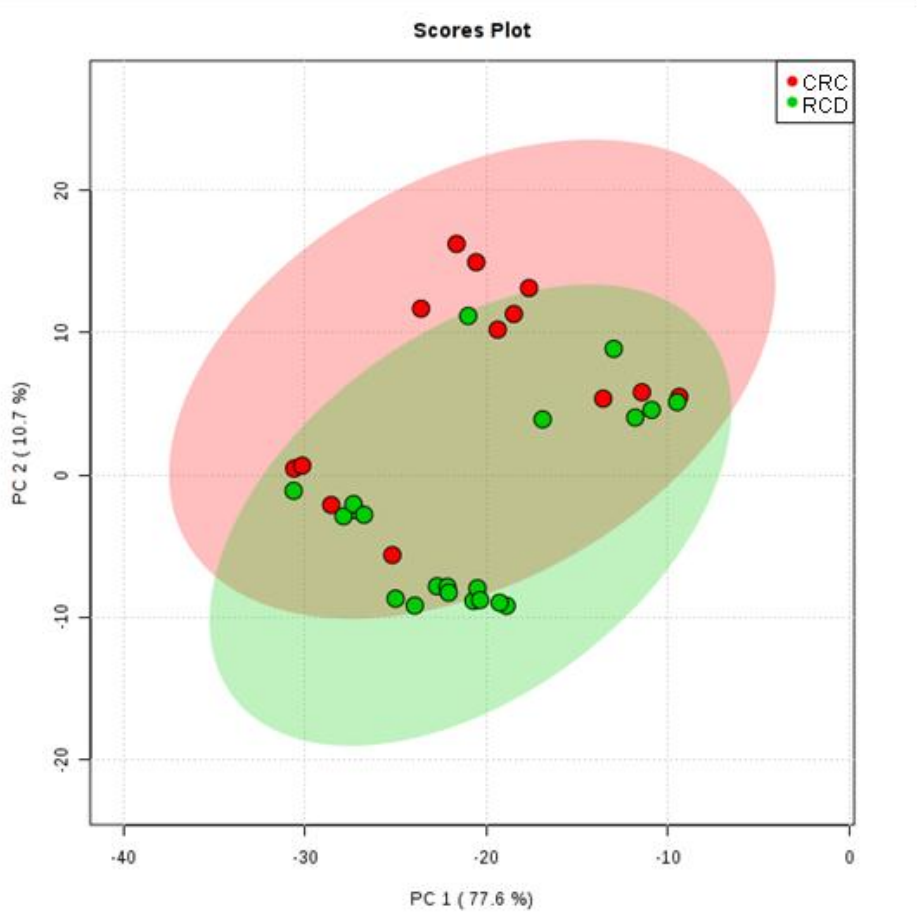

Figure 6.1: PCA score plot of the verified and top ranked features of the respective platforms. Separation of RCD patient group (green) and CRC group (red) when using all the features listed in Table 6.1.

Figure 6.1 shows the PCA score plot of the RCD patient group and CRC group when using the verified and top ranked features of the respective platforms listed in Table 6.1. In Figure 6.1, the RCD group and CRC group are scattered across the plot, with no clustering of groups and no separation between these two groups. However, it should be noted that PCA is merely an unsupervised visualization technique which could give a different result when more samples are included. Therefore, clinicians tend to make greater use of ROC curves (with simple classification algorithms) to classify cases (Xia *et al*., 2013).

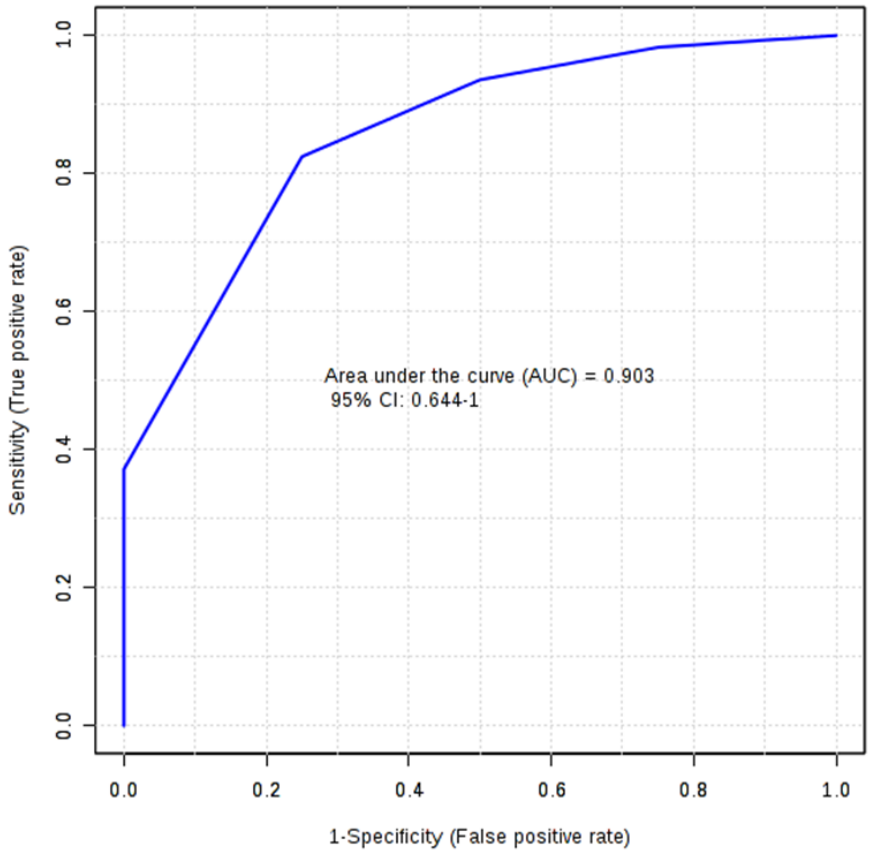

Figure 6.2: ROC curve of the verified and top ranked features of respective platforms. The curve displays the evaluation results of the verified and top ranked features of respective platforms with regard to sensitivity and specificity.

The discrimination power of the biosignature is shown in Figure 6.2. This ROC curve was constructed using the verified LC-MS features and the top ranked features of the GC-MS and NMR analyses. The AUC of 0.903 with a 95% confidence interval that ranges between 0.644 and 1 indicates that this signature is both specific and sensitive and can be regarded as acceptable.

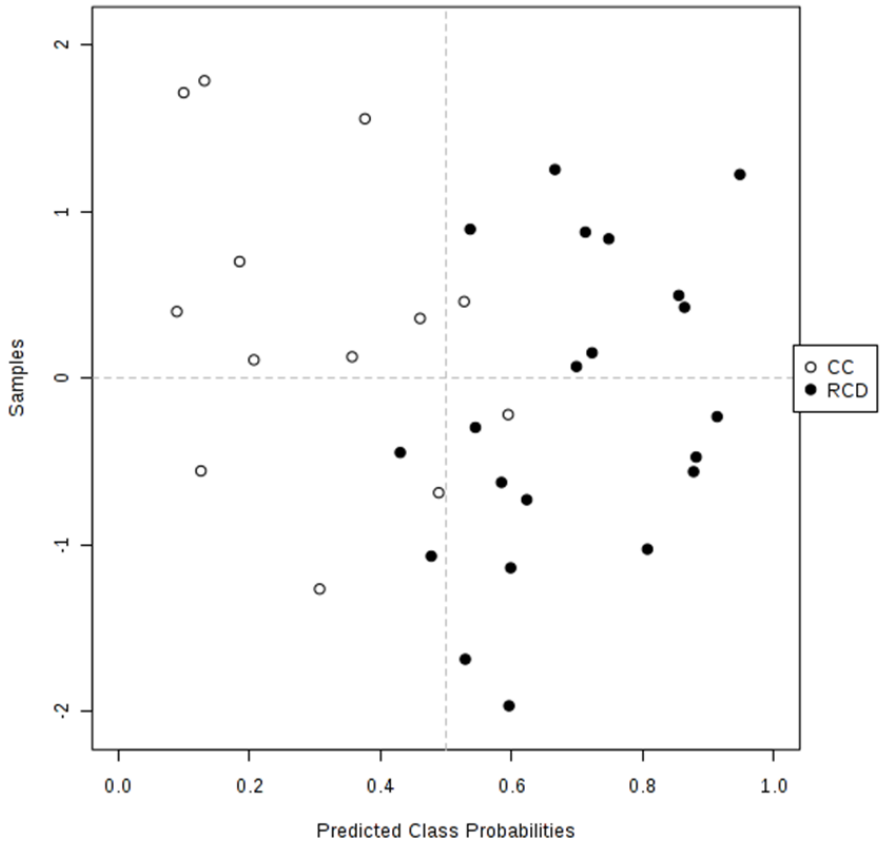

Figure 6.3: Average predicted class probability of the experimental groups over 100 cross validations using the verified and top ranked features of the respective platforms. The predicated class probabilities for each sample using all features listed in Table 6.1. The black dots (●) represent the RCD patient group, while the white dots (○) represent the CRC group.

Figure 6.3 shows the average predicted class probability of each sample across 100 validations. A balanced sub-sampling approach was used by the validation algorithm with a classification boundary in the centre (0.5). On the left side of this figure the CRC sample group can be seen and to the right side the samples classified as RCDs are visible. As can be seen in the figure, some of the samples are not correctly classified. While this figure is merely an indication of the performance of the model, it should be mentioned that the classification boundary can be modified (in conjunction with the clinician) to include all true positive cases in clinical application (for example 0.4).

6.3 Improved biosignature feature selection

Even though the combined list of 14 features had good discriminative power, a prerequisite of a biosignature is that the signature must be the smallest number of metabolites that have the best discriminative power between the experimental groups (Xia *et al.,* 2013). For this reason, statistical analyses were used to firstly indicate the number and combination of features that should be used to form part of the improved biosignature. Secondly, the improved biosignature had to be evaluated in order to establish the discriminative power of this improved biosignature.

The discriminative power of the evaluated signatures using multivariate statistical analyses is shown in ROC curve format in Figure 6.4. The six graphs displayed in Figure 6.4 each represent a sensitivity/specificity value for a particular decision threshold. Graphs representing 13, 10, 7, 5, 3 and 2 features are visible.

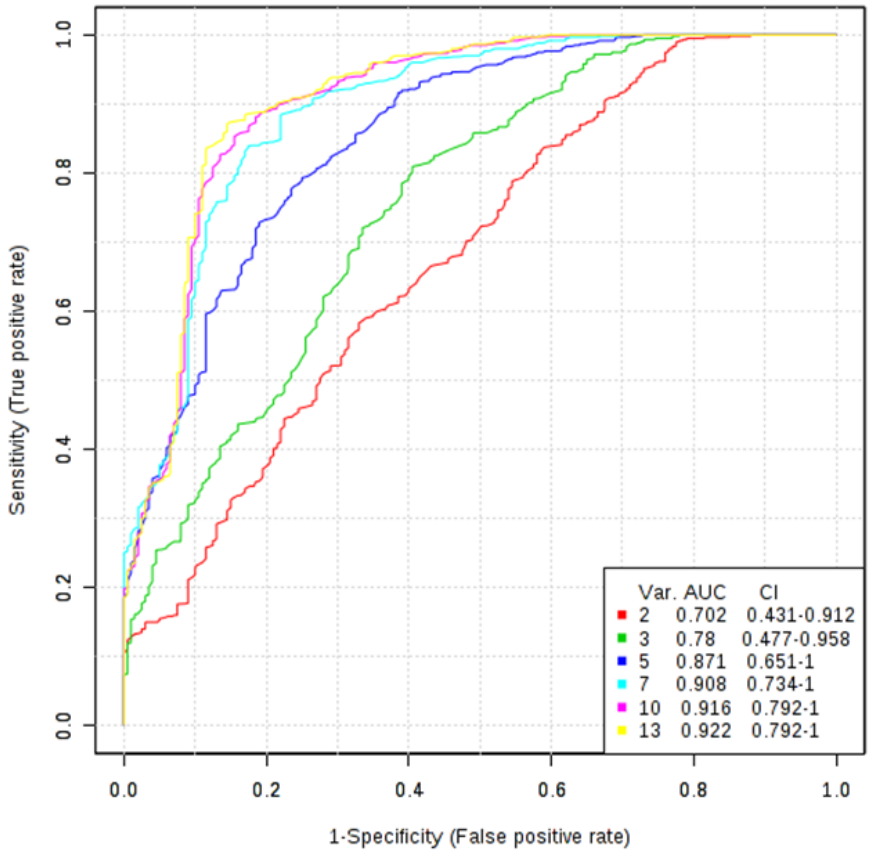

Figure 6.4: Multivariate AUC of the verified and top ranked features. In this figure the confidence and performance intervals of a different number of selected variables are given. Var: Variables, CI: Confidence interval and AUC: area under the curve.

Results indicated that the greater the number of variables used, the better the model performed. The decision was made to use five variables to form the improved biosignature due to the fact that the least amount of features possible should be used in a biosignature. The five variable model indicated an acceptable AUC of 0.871. Multivariate ROC (with PLS-DA as the underlying algorithm) was used to determine the best possible combination of features to include in the improved biosignature. Figure 6.5 indicates the important features, as identified by the PLS-DA algorithm.

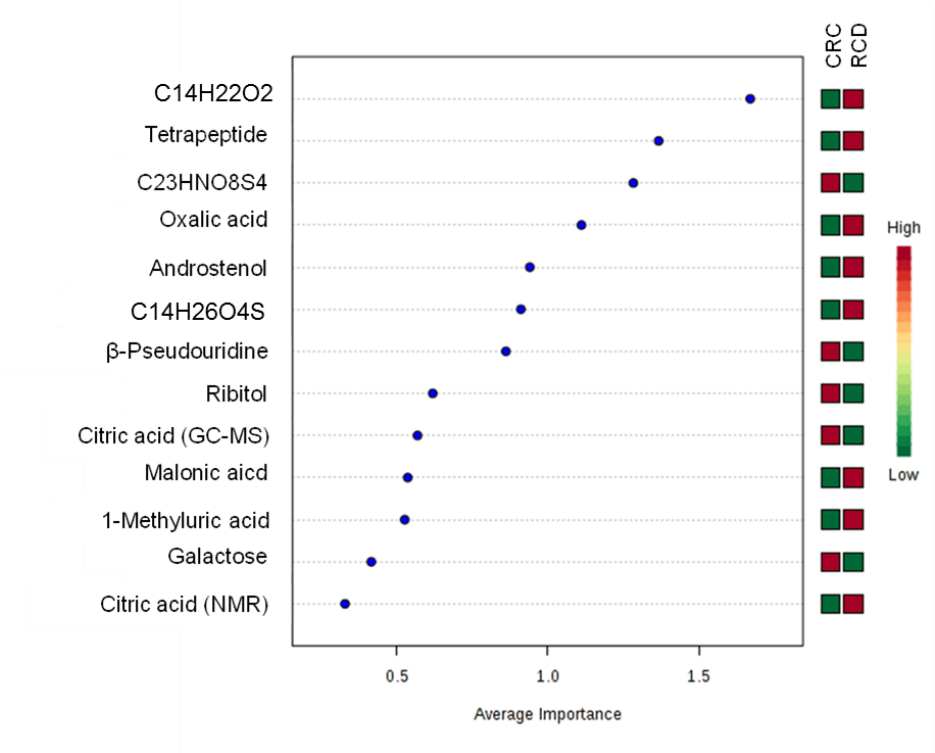

Figure 6.5: Important features identified by PLS-DA. The coloured boxes to the right indicate the relative concentrations of the metabolite in the RCD and CRC groups.

As mentioned previously, the ROC curve indicated that the five variable model performed well and could be used to construct the improved biosignature. In Figure 6.5 the top five features were included in the improved biosignaure, since the statistical test used indicated that these features could be regarded as important. Table 6.2 lists the features included in the improved biosignature and the platform from which it originated.

Table 6.2: The improved biosignature

Annotated name	Platform
1. $C_{14}H_{22}O_2$	LC-QTOF
2. Tetrapeptide	LC-QTOF
3. $C_{23}HNO_8S_4$	LC-QTOF
4. Oxalic acid	GC-MS
5. Androstenol	LC-QTOF

In Table 6.2 the improved biosignature is shown. The improved biosignature consisted of four features of the LC-QTOF and one feature from GC-MS. Based on statistical test results, none of the top ranked features of the NMR analyses contributed to the improved biosignature. Oxalic acid and androstenol (which is a steroid compound) have not previously been described for RCDs and the presence of these metabolites in the urine of this sample cohort is uncertain. Venter *et al*. (2015) speculated that the features of the proposed biosignature might be di- and tri-peptides. The tetrapeptide may be due to increased protein degradation found in RCD patients. The two remaining features (Features 1 and 3, as listed in Table 6.2) could not be identified using the databases at our disposal. It is however once again important to mention that the names given are only potential identifications and still need to be validated using reference standards. For this reason, further investigations are required to confirm identifications in order to provide a clear biological interpretation of this feature list.

6.4 Discriminative power of the improved biosignature

Following statistical analyses and the selection of the improved biosignature, the discriminative power of the improved biosignature had to be evaluated in order to determine whether or not the combination of the selected features had the ability to distinguish between these two experimental groups. For evaluation of the discriminative power a PCA score plot and a ROC curve were used.

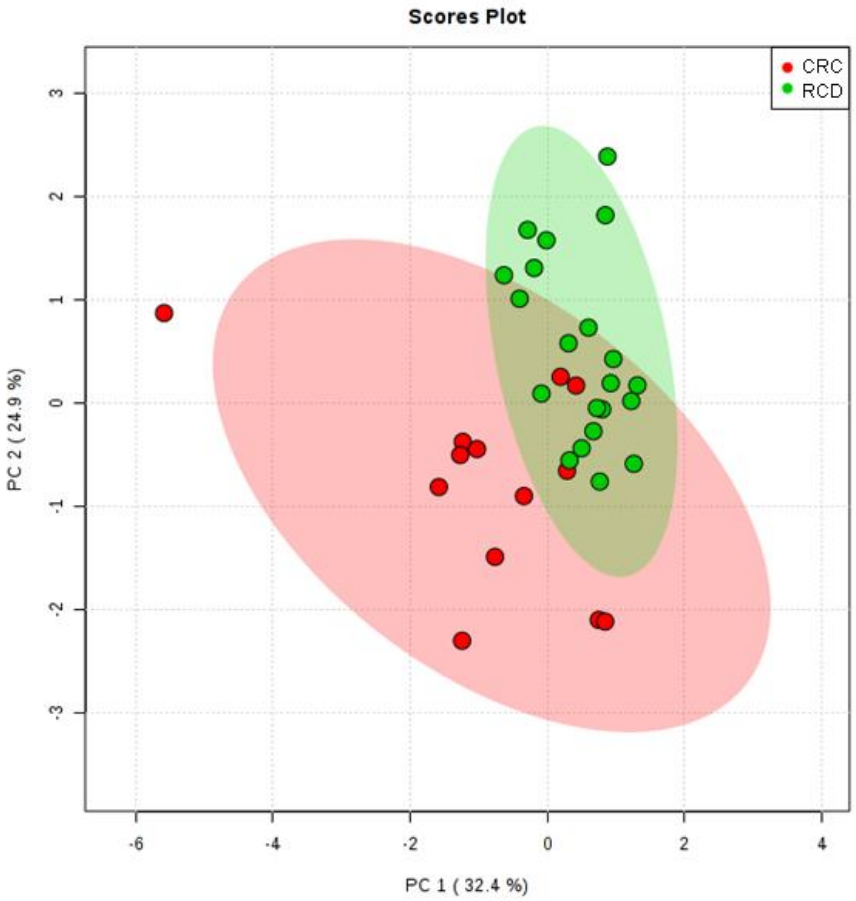

Figure 6.6: PCA score plot of the improved biosignature. The figure displays the discriminative power of the RCD patient group and CRC group with the use of the improved biosignature.

Figure 6.6 shows the PCA score plot of the RCD patient group and CRC group when using the improved biosignature listed in Table 6.2. In Figure 6.6 the RCD group is represented by the smaller circled group in green and the larger circled group in red represents the CRC group. With the use of the 95% confidence region, three of the CRC samples were included within the RCD group region while ten were not included. If the PCA was used for classification, then three CRC patients would have been included in the RCD group receiving a muscle biopsy (to confirm the absence or presence of an RCD). As mentioned previously, the ultimate goal would be to obtain complete separation between these groups. However, complete separation is not necessary, as long as the number of false positive cases included for biopsy can be reduced. When taking into account that selected CRC patients had enzyme activity scores very close to RCD positive scores, and thus metabolic profiles similar to those of RCD patients, it indicates that it is thus impossible to create a model where these groups separate perfectly. This could also explain why some of the CRC patients clustered with the RCD patients.

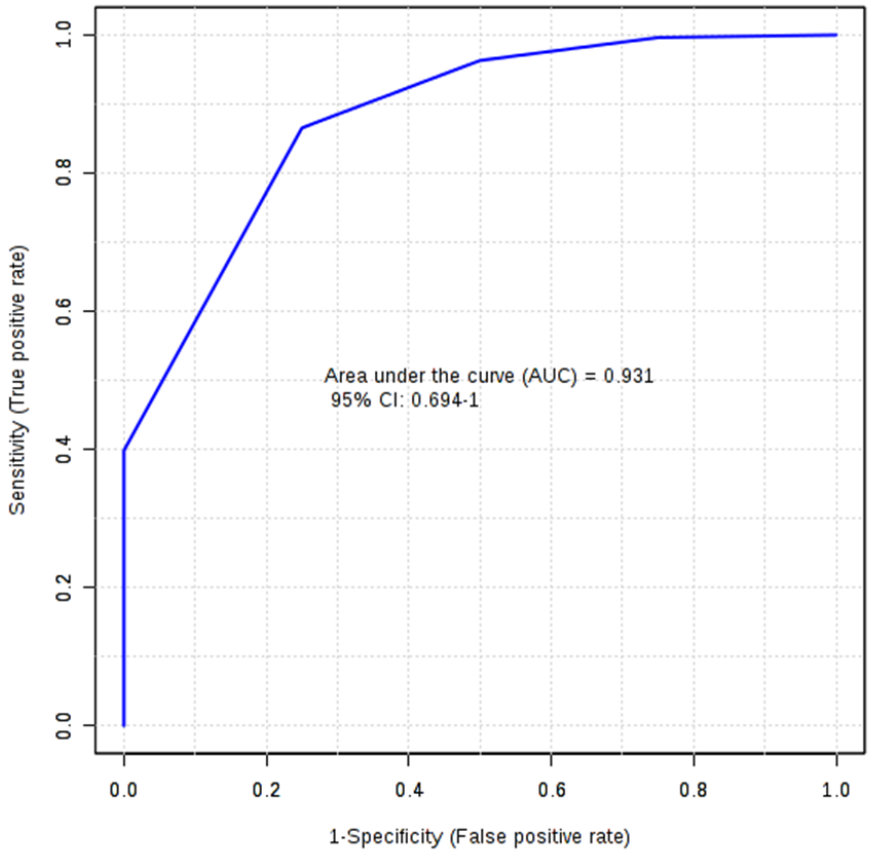

Figure 6.7: ROC curve for the improved biosignature. The curve displays the results of the evaluation of the improved biosignature with regard to sensitivity and specificity.

The ROC curve of the improved biosignature is shown Figure 6.7. The AUC of 0.931 with a 95% confidence interval that ranged between 0.694 and 1 indicated that this signature was both specific and sensitive and could be regarded as very good. When comparing the ROC curve (Figure 6.2) of the 14 verified and top ranked features (listed in Table 6.1) and the ROC curve of the improved biosignature in Figure 6.7, it can be seen that the improved biosignature performed better.

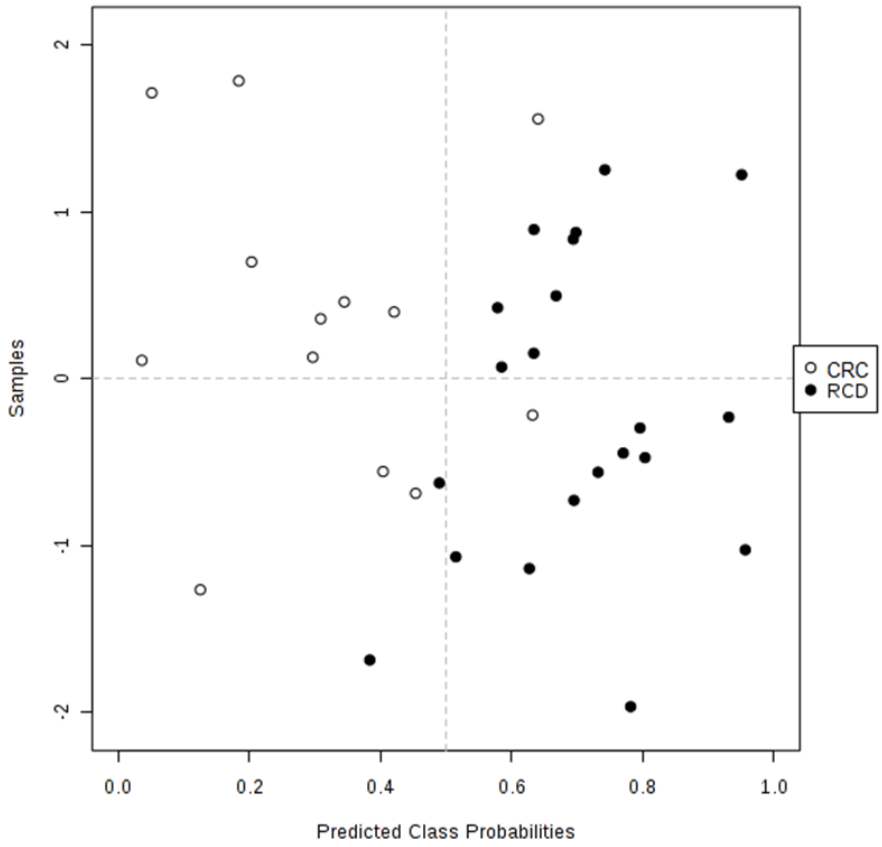

Figure 6.8: Average predicted class probability of the experimental groups over 100 cross validations using the verified and top ranked features of the respective platforms. The predicated class probabilities for each sample using all features listed in Table 6.1. The black dots (●) represent the RCD patient group, while the white dots (○) represent the CRC group.

Figure 6.4 shows the average predicted class probability of each sample across 100 validations. The average predicted class probability of samples across 100 cross validations were calculated by a balanced sub-sampling approach with the classification boundary at 0.5. Results indicated that the biosignature was able to classify the samples accurately in the 100 cross validation sub-samples of the original data with the exception of four samples. In Figure 6.4 it can be seen that two of the CRC samples were classified as RCDs, which means that these patients will be sent for a biopsy as well. Two RCD samples were classified as CRC, which means that when the decision boundary is set at 0.5 these patients will missed in the screening process. Important to keep in mind that the decision boundary can be adjusted in this case to 0.4 which means that all the RCD would be included as well as some of the CRC control samples. Even though more CRC control samples will be included, the improve biosignature still reduces the number of false positive included for biopsy.

In order to validate performance of the improved biosignature, permutation testing was used. This technique was used as another form of cross-validation and is based on the hypothesis that the biosignature could have been found a second time if every control or patients sample was randomly allocated to an alternative group (Xia *et al*., 2013). To test the hypothesis, 1 000 models were permutated by randomly allocating samples to one of the experimental groups and then applying the improved biosignature for classification of the groups. The reference distribution is compared to the correctly assigned model by generating a p-value. When this p-value is < 0.05, it indicates that the randomly permutated outcome variable has less than 5% chance of generating a model that will perform similar to that of the correctly assigned model (Xia *et al.,* 2013). The results of the improved biosignature resulted in a permutation test value of $p < 0.02$, thus by using random guessing there is a 2% chance of producing the same model, thus indicating the specificity of the biosignature. Lastly, it could be verified that after using 1 000 permutation tests, none of the results performed better than the original design, therefore a p-value of $p < 0.02$ was generated.

6.5 Summary

In this chapter the verified features of the proposed biosignature, together with the top ranked features of the GC-MS and NMR analysis were used to compile a list of features that were considered for the improved biosignature. Prior to final feature selection of the improved biosignature, the discriminative power of the 15 features were evaluated. The results indicated that the combined list of 15 features had sufficient discriminative power. However, a prerequisite of a biosignature is that the signature should contain the least number of metabolites with the best discriminative power. Following statistical analyses, the five feature model proved to be sufficient. Multivariate ROC (with PLS-DA as the underlying algorithm) was used to select the best combination of five features to form the improved biosignature. The discriminative power of the improved biosignature was evaluated and proved to be sufficient for distinguishing between the two experimental groups.

CHAPTER 7

CONCLUSION

7

7.1 Introduction

RCDs are the largest sub-group of mitochondrial diseases and are considered to be one of the most common forms of inherited metabolic diseases, with an estimated incidence of one in every 5 000 - 10 000 live births. Diagnosis of RCDs is complicated due to heterogeneity and requires a multi-disciplinary approach. Currently, enzyme analyses performed on a muscle sample obtained from a muscle biopsy are considered the golden standard for diagnosing RCDs. However, a muscle biopsy is a highly invasive procedure with many limitations. Another challenge that clinicians face, is the selection of patients who must undergo a muscle biopsy. For this reason, other less invasive options ought to be explored to aid clinicians in patient selection. Venter *et al*. (2015) proposed a biosignature for RCDs consisting of 12 features, but where only the accurate mass and retention time of the features was reported. Since the true identity of the 12 features was unknown, the biosignature had to be verified using the limited information available. This biosignature also had the potential to be expanded, in terms of metabolome coverage, by performing untargeted NMR and GC-MS analyses on the identical sample cohort in order to identify discriminative features that could form part of the improved biosignature, potentially missed with the LC-QTOF analysis Venter *et al.* (2015) used to construct the proposed biosignature.

In this Chapter the main findings of the study will be discussed linked to the specific objectives set for the study in order to achieve the aim of improving the biosignature proposed by Venter *et al.* (2015).

7.2 Conclusions

7.2.1 Objective 1 – First level of verification

The first level of verification was part of the first objective and in order to achieve this objective, two goals had to be reached. Using the criteria mentioned in Section 3.5, a total of five features were verified and were included in the improved biosignature. False features are a common occurrence in metabolomics data, with these artefacts being a result of chemical or bioinformatics noise (Chong et al., 2015). The artefacts of analysis (or data processing) and the features that could not be detected again were eliminated, resulting in a list of verified features that had little discriminative power when combined. Even though the combination of these five features had less discriminative power than the proposed biosignature of 12 features, these five features could be detected on multiple instruments and using an alternative software program, therefore making them true features.

7.2.2 Objective 2 – Second level of verification

The second level of verification involved characterization of features in an attempt to gain more information and was part of the second objective. Metabolite identification proved to be the most difficult part of this study and is regarded in literature as a large bottleneck in metabolomics studies. Databases are essential for metabolite identification and are therefore one of the main limitations due to the high number of metabolite entries and limited experimental fragmentation spectra that is used for spectral matching (Alonso et al., 2015). For instance databases such as Metlin and HMDB contain ~241 000 and ~42 000 metabolite entries, respectively, only 13 000 of these have experimental fragmentation spectra that could be used for spectral matching to make identifications. Since no conclusive fragmentation spectra could be generated it was impossible to perform spectral matching. One possible reason as to why quality fragmentation spectra could not be obtained was that the concentrations of the features in the biological sample tested may have been too low. The features of the proposed biosignature are not necessarily the features in the biological sample that have the highest concentration. The concentrations of features varied between patients, therefore when the pooled urine sample was prepared, the feature concentration was diluted thus resulting in feature concentrations that were either below the detection limit or that were too low to generate quality fragmentation spectra fit for spectral matching. The theoretical approach in silico fragmentation yielded only one possible identification, this identification still has to be verified using reference standards.

7.2.3 Objective 3 – GC-MS analysis and identification of discriminative features

As part of the third objective for achieving the aim of this study, the identical sample cohort was analysed using the evaluated GC-MS method in order to expand the proposed biosignature. Following statistical analyses average rank of each feature across all the tests was determined and an average ranking of important features graph could be constructed. The average ranking graph was used as a guideline to select the top six features (Table 4.3) that was included in the improved biosignature. The discriminative power of the top ranked features was evaluated. Results indicated that this combination of features had no discriminative power. The reason for this may be that the when using GC-MS analyses the primary metabolism is focused on, these two groups may have similar primary metabolism metabolite profiles resulting in no variation that could distinguish between these two groups.

7.2.4 Objective 4 – NMR analysis and identification of discriminative features

Following statistical analyses average rank of each feature across all the tests was determined and an average ranking of important features graph could be constructed. Using the average rank graph, the top four features (Table 5.3) were selected that formed part of the proposed biosignature. The discriminative power of the top ranked features was evaluated. Results indicated that this combination of features had no discriminative power. Sensitivity of NMR remains a limitation of this technique, with low concentration features masked by high concentration features. As mentioned previously the important features for biological variance aren't necessarily the features with the highest concentration resulting features that could have had high discriminative power being missed.

7.2.5 Objective 5 - Compilation of an improved biosignature and evaluation of discriminative power

The fifth objective was compiling a list of features that would form part of the improved biosignature and then evaluated the discriminative power of this signature. The five verified and ten top ranked features of the respective platforms were combined and the discriminative power of these features together was evaluated. The PCA score plot indicated that the combination of features were not able to distinguish between these two groups, although the ROC curve and average predicted class probability indicated that combination of features performed good in classifying these two groups.

Even though the combined list of 15 features had good discriminative power a prerequisite of a biosignature is the signature must be the smallest number of metabolites with the best discriminative power (Xia *et al*., 2013). Therefore statistical analyses were used to find the best combination of five features that would form part of the improved biosignature. The improved biosignature consisted of four LC-MS features and one GC-MS feature. The limited number of features that contributed to the improved biosignature by the additional platforms may be due to a number of reasons. For GC-MS analyses focus is generally placed on the primary metabolism. The two groups investigated may have had similar primary metabolism metabolite profiles, thus resulting in a lack of variation to distinguish between these two groups; this may be the reason why only one feature contributed to the improved biosignature. For NMR sensitivity remains a limitation of this technique, with low concentration metabolites masked by high concentration metabolites. As mentioned previously the important features for biological variance aren't necessarily the features with the highest concentration resulting in features that could have had high discriminative power being missed. This may have been the reason why NMR did not contributed a single metabolite to the improved biosignature. When comparing this result to that of Smuts *et al*. (2013), where only one metabolite (creatine) were part of the biosignature constructed to distinguish between healthy controls and RCD patients, it is likely the case that the sensitivity of NMR remains the limiting factor.

Following the evaluation of the discriminative power, the results indicated that the improved biosignature is both specific and sensitive and can be regarded as good. Even though the proposed biosignature of Venter *et al*. (2015) could classify these two groups 100% accurately, it contained false features that could not be verified. Additionally the proposed biosignature consists of features that are detected using two different analyses (different LC columns and ionization modes) which is time consuming. The improved biosignature consists of four verified LC-MS features and a GC-MS feature that could be analysed in a single LC-MS method analyses. Even though the improved biosignature is unable to classify samples 100% accurately with some of the CRC samples classified as RCDs, it can still be helpful in limiting the inclusion of CRC patients in the biopsy process which gives it the potential to be used in the diagnostic workflow.

7.3 Final conclusion

The proposed biosignature of Venter *et al*. (2015) was improved in two ways. Firstly, the features that may be artefacts of analysis or data processing were eliminated, resulting in five features being verified. Secondly, even though the degree of separation between these two groups are less than that of the original proposed biosignature, the improved biosignature consists of fewer features (five instead of twelve) and could still distinguish between these two groups to a certain extent. Additionally, the compounds of the improved biosignature have the potential to be analysed using a LC-MS method, where all the compounds can be measured using a single analysis instead of two, as was the case with the proposed biosignature where a positive and negative ionisation methods with two different columns had to be performed.

The ultimate goal would be to have complete separation between these two groups. As mentioned previously, it is not necessarily possible or needed. The aim of this study was to improve the existing biosignature in two phases by verifying and expanding the proposed biosignature of Venter *et al*. (2015). Five of the features of the proposed biosignature were verified, of which four were included in the improved biosignature. For the expanding phase, following statistical analyses only one additional feature of GC-MS analyses was included in the improved biosignature.

7.4 Future recommendations

The identity of the features remains a problem. In hindsight, it is recommended that rather than using a pooled urine sample, a sample with the highest concentration of the specific feature should be used in analyses and fraction collection. This may be helpful in concentrating the sample in order to generate quality fragmentation spectra.

Before this biosignature can be implemented in a clinical environment a number of verification steps still need to be performed. Firstly the oxalic acid metabolite identified as an important feature with the GC-MS analysis still need to be verified on the first level of verification. Secondly the next step for this improved biosignature is the third and most important level of verification. This level involves the analysis of a different cohort of samples and to determine whether or not this biosignature is able to distinguish between the RCD patient group and CRC group. A study like this would greatly benefit from more homogenous experimental groups. This would exclude influences such as age, gender, ethnicity and additional factors that may influence the results obtained.

Four compounds of the improved biosignature originated from the LC-QTOF, while the oxalic acid was detected using the GC-MS. It is however possible to measure oxalic acid using a LC-MS method. It is therefore recommended to develop a standardized LC-MS method in order to measure the five compounds of the improved biosignature in a single analysis. Measuring the improved biosignature in a single analysis will save time and cost per sample analysis which will be ideal when implementing this biosignature in a clinical environment.

REFERENCES

ABDI, H. & WILLIAMS, L. J. 2010. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2, 433-459.

ALONSO, A., MARSAL, S. & JULIÀ, A. 2015. Analytical methods in untargeted metabolomics: state of the art in 2015*. Frontiers in bioengineering and biotechnology*, 3, 23.

ÁLVAREZ-SÁNCHEZ, B., PRIEGO-CAPOTE, F. & DE CASTRO, M. L. 2010a. Metabolomics analysis I. Selection of biological samples and practical aspects preceding sample preparation. *TrAC Trends in Analytical Chemistry*, 29, 111-119.

ÁLVAREZ-SÁNCHEZ, B., PRIEGO-CAPOTE, F. & DE CASTRO, M. L. 2010b. Metabolomics analysis II. Preparation of biological samples prior to detection. *TrAC Trends in Analytical Chemistry*, 29, 120-127.

APPLEGARTH, D. A. & TOONE, J. R. 2000. Incidence of inborn errors of metabolism in British Columbia, 1969–1996. *Pediatrics*, 105, e10-e10.

BANDYOPADHYAY, S. & DUTTA, A. 2010. Mitochondrial Medicine. *Journal of the association of physicians of India*, 58, 237.

BARSHOP, B. A. 2004. Metabolomic approaches to mitochondrial disease: correlation of urine organic acids. *Mitochondrion*, 4, 521-527.

BERG, J. M., TYMOCZKO, J. L. & STRYER, L. 2002a. Carbon atoms of degraded amino acids emerge as major metabolic intermediates. *Biochemistry*. 5th edition. New York: W H Freeman

BERG, J. M., TYMOCZKO, J. L. & STRYER, L. 2002b. The citric acid cycle. *Biochemistry*. 5th edition. New York: W H Freeman

BOLISETTY, S. & JAIMES, E. A. 2013. Mitochondria and reactive oxygen species: physiology and pathophysiology. *International journal of molecular sciences*, 14, 6306-6344.

BUTLER, J. A., MISHUR, R. J., BHASKARAN, S. & REA, S. L. 2013. A metabolic signature for long life in the Caenorhabditis elegans Mit mutants. *Aging cell*, 12, 130-138.

BYKHOVSKAYA, Y., CASAS, K., MENGESHA, E., INBAL, A. & FISCHEL-GHODSIAN, N. 2004. Missense mutation in pseudouridine synthase 1 (PUS1) causes mitochondrial myopathy and sideroblastic anemia (MLASA). *The American Journal of Human Genetics*, 74, 1303-1308.

CHERNUSHEVICH, I. V., LOBODA, A. V. & THOMSON, B. A. 2001. An introduction to quadrupole–time-of-flight mass spectrometry. *Journal of Mass Spectrometry*, 36, 849-865.

CHONG, E. Y., HUANG, Y., WU, H., GHASEMZADEH, N., UPPAL, K., QUYYUMI, A. A., JONES, D. P. & YU, T. 2015. Local false discovery rate estimation using feature reliability in LC/MS metabolomics data. *Scientific reports*, 5.

CHRISTIN, C., HOEFSLOOT, H. C., SMILDE, A. K., HOEKMAN, B., SUITS, F., BISCHOFF, R. & HORVATOVICH, P. 2013. A critical assessment of feature selection methods for biomarker discovery in clinical proteomics. *Molecular & Cellular Proteomics*, 12, 263-276.

CLARKE, N. J., RINDGEN, D., KORFMACHER, W. A. & COX, K. A. 2001. Peer reviewed: Systematic LC/MS metabolite identification in drug discovery. *Analytical chemistry*, 73, 430 A-439 A.

COHEN, B. H. & GOLD, D. R. 2001. Mitochondrial cytopathy in adults: what we know so far. *Cleveland Clinic journal of medicine*, 68, 625-642.

DARIN, N., OLDFORS, A., MOSLEMI, A. R., HOLME, E. & TULINIUS, M. 2001. The incidence of mitochondrial encephalomyopathies in childhood: clinical features and morphological, biochemical, and DNA abnormalities. *Annals of neurology*, 49, 377-383.

DEJAEGHER, B. & VANDER HEYDEN, Y. 2010. HILIC methods in pharmaceutical analysis. *Journal of separation science*, 33, 698-715.

DETTMER, K., ARONOV, P. A. & HAMMOCK, B. D. 2007. Mass spectrometry-based metabolomics. *Mass spectrometry reviews*, 26, 51-78.

DIMAURO, S. 2004. Mitochondrial diseases. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, 1658, 80-88.

DIMAURO, S. & SCHON, E. A. 2003. Mitochondrial respiratory-chain diseases. *New England Journal of Medicine*, 348, 2656-2668.

DONA, A. C., JIMÉNEZ, B., SCHÄFER, H., HUMPFER, E., SPRAUL, M., LEWIS, M. R., PEARCE, J. T., HOLMES, E., LINDON, J. C. & NICHOLSON, J. K. 2014. Precision high-throughput proton nmr spectroscopy of human urine, serum, and plasma for large-scale metabolic phenotyping. *Analytical chemistry*, 86, 9887-9894.

DUNN, W. B. 2008. Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes. *Physical biology*, 5, 011001.

DUNN, W. B., BAILEY, N. J. & JOHNSON, H. E. 2005. Measuring the metabolome: current analytical technologies. *Analyst*, 130, 606-625.

DUNN, W. B., BROADHURST, D., BEGLEY, P., ZELENA, E., FRANCIS-MCINTYRE, S., ANDERSON, N., BROWN, M., KNOWLES, J. D., HALSALL, A. & HASELDEN, J. N. 2011a. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nature protocols*, 6, 1060-1083.

DUNN, W. B., BROADHURST, D. I., ATHERTON, H. J., GOODACRE, R. & GRIFFIN, J. L. 2011b. Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy. *Chemical Society Reviews*, 40, 387-426.

DUNN, W. B. & ELLIS, D. I. 2005. Metabolomics: current analytical platforms and methodologies. *TrAC Trends in Analytical Chemistry*, 24, 285-294.

DUNN, W. B., ERBAN, A., WEBER, R. J., CREEK, D. J., BROWN, M., BREITLING, R., HANKEMEIER, T., GOODACRE, R., NEUMANN, S. & KOPKA, J. 2013. Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics*, 9, 44-66.

ELLINGER, J. J., CHYLLA, R. A., ULRICH, E. L. & MARKLEY, J. L. 2013. Databases and software for NMR-based metabolomics. *Current Metabolomics*, 1.

EMWAS, A.-H. M., SALEK, R. M., GRIFFIN, J. L. & MERZABAN, J. 2013. NMR-based metabolomics in human disease diagnosis: applications, limitations, and recommendations. *Metabolomics*, 9, 1048-1072.

ESTEITIE, N., HINTTALA, R., WIBOM, R., NILSSON, H., HANCE, N., NAESS, K., TEÄR-FAHNEHJELM, K., VON DÖBELN, U., MAJAMAA, K. & LARSSON, N. G. 2005. Secondary metabolic effects in complex I deficiency. *Annals of neurology*, 58, 544-552.

FAGES, A., PONTOIZEAU, C., JOBARD, E., LÉVY, P., BARTOSCH, B. & ELENA-HERRMANN, B. 2013. Batch profiling calibration for robust NMR metabonomic data analysis. *Analytical and bioanalytical chemistry*, 405, 8819-8827.

FAHY, E., COTTER, D., SUD, M. & SUBRAMANIAM, S. 2011. Lipid classification, structures and tools. *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids*, 1811, 637-647.

FALK, M., ZHANG, Z., ROSENJACK, J., NISSIM, I., DAIKHIN, E., SEDENSKY, M., YUDKOFF, M. & MORGAN, P. 2008. Metabolic pathway profiling of mitochondrial respiratory chain mutants in C. elegans. *Molecular genetics and metabolism*, 93, 388-397.

FERNÁNDEZ-PERALBO, M. & DE CASTRO, M. L. 2012. Preparation of urine samples prior to targeted or untargeted metabolomics mass-spectrometry analysis. *TrAC Trends in Analytical Chemistry*, 41, 75-85.

FERNÁNDEZ-VIZARRA, E., TIRANTI, V. & ZEVIANI, M. 2009. Assembly of the oxidative phosphorylation system in humans: what we have learned by studying its defects. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1793, 200-211.

FERNIE, A. R., CARRARI, F. & SWEETLOVE, L. J. 2004. Respiratory metabolism: glycolysis, the TCA cycle and mitochondrial electron transport. *Current opinion in plant biology*, 7, 254-261.

FIEHN, O., KOPKA, J., TRETHEWEY, R. N. & WILLMITZER, L. 2000. Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. *Analytical chemistry*, 72, 3573-3580.

GARRETT, R. H. & GRISHAM, C. M. 2008. Metabolism: An Overview. *Biochemistry*. 4th Edition ed.: *Brookes*-Cole.

GOLDSMITH, P., FENTON, H., MORRIS-STIFF, G., AHMAD, N., FISHER, J. & PRASAD, K. R. 2010. Metabonomics: a useful tool for the future surgeon. *Journal of Surgical Research*, 160, 122-132.

GORMAN, G. S., SCHAEFER, A. M., NG, Y., GOMEZ, N., BLAKELY, E. L., ALSTON, C. L., FEENEY, C., HORVATH, R., YU-WAI-MAN, P. & CHINNERY, P. F. 2015. Prevalence of nuclear and mitochondrial DNA mutations related to adult mitochondrial disease. *Annals of neurology*, 77, 753-759.

CRAIG, A., CLOAREC, O., HOLMES, E., NICHOLSON, J. K. & LINDON, J. C. 2006. Scaling and normalization effects in NMR spectroscopic metabonomic data sets. *Analytical chemistry*, 78, 2262-2267.

GREAVES, L. C., REEVE, A. K., TAYLOR, R. W. & TURNBULL, D. M. 2012. Mitochondrial DNA and disease. *The Journal of pathology*, 226, 274-286.

GULLBERG, J., JONSSON, P., NORDSTRÖM, A., SJÖSTRÖM, M. & MORITZ, T. 2004. Design of experiments: an efficient strategy to identify factors influencing extraction and derivatization of Arabidopsis thaliana samples in metabolomic studies with gas chromatography/mass spectrometry. *Analytical biochemistry*, 331, 283-295.

HAAS, R. H., PARIKH, S., FALK, M. J., SANETO, R. P., WOLF, N. I., DARIN, N. & COHEN, B. H. 2007. Mitochondrial disease: a practical approach for primary care physicians. *Pediatrics*, 120, 1326-1333.

HAAS, R. H., PARIKH, S., FALK, M. J., SANETO, R. P., WOLF, N. I., DARIN, N., WONG, L.-J., COHEN, B. H. & NAVIAUX, R. K. 2008. The in-depth evaluation of suspected mitochondrial disease. *Molecular genetics and metabolism*, 94, 16-37.

HERRMANN, J. M. & NEUPERT, W. 2000. Protein transport into mitochondria. *Current opinion in microbiology*, 3, 210-214.

HRYDZIUSZKO, O. & VIANT, M. R. 2012. Missing values in mass spectrometry based metabolomics: an undervalued step in the data processing pipeline. *Metabolomics*, 8, 161-174.

IVANISEVIC, J., BENTON, H. P., RINEHART, D., EPSTEIN, A., KURCZY, M. E., BOSKA, M. D., GENDELMAN, H. E. & SIUZDAK, G. 2014. An interactive cluster heat map to visualize and explore multidimensional metabolomic data. *Metabolomics*, 1-6.

KAMI, K., FUJITA, Y., IGARASHI, S., KOIKE, S., SUGAWARA, S., IKEDA, S., SATO, N., ITO, M., TANAKA, M. & TOMITA, M. 2012. Metabolomic profiling rationalized pyruvate efficacy in cybrid cells harboring MELAS mitochondrial DNA mutations. *Mitochondrion*, 12, 644-653.

KOEK, M. M., MUILWIJK, B., VAN DER WERF, M. J. & HANKEMEIER, T. 2006. Microbial metabolomics with gas chromatography/mass spectrometry. *Analytical chemistry*, 78, 1272-1281.

KOENE, S. & SMEITINK, J. 2011. Mitochondrial medicine. *Journal of inherited metabolic disease*, 34, 247-248.

KOENIG, M. K. 2008. Presentation and diagnosis of mitochondrial disorders in children. *Pediatric neurology*, 38, 305-313.

KRAUSS, S. 2001. Mitochondria: Structure and role in respiration. Beth Israel Deaconess Medical Center and Harvard Medical School, USA. Encyclopedia of Life Sciences. *Nature Publishing Group*, p.1-6.

LEONARD, J. & SCHAPIRA, A. H. 2000. Mitochondrial respiratory chain disorders I: mitochondrial DNA defects. *The Lancet*, 355, 299-304.

LEONG, D. W., KOMEN, J. C., HEWITT, C. A., ARNAUD, E., MCKENZIE, M., PHIPSON, B., BAHLO, M., LASKOWSKI, A., KINKEL, S. A. & DAVEY, G. M. 2012. Proteomic and metabolomic analyses of mitochondrial complex I-deficient mouse model generated by spontaneous B2 short interspersed nuclear element (SINE) insertion into NADH dehydrogenase (ubiquinone) Fe-S protein 4 (Ndufs4) gene. *Journal of Biological Chemistry*, 287, 20652-20663.

LIANG, C., AHMAD, K. & SUE, C. M. 2014. The broadening spectrum of mitochondrial disease: Shifts in the diagnostic paradigm. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1840, 1360-1367.

LILAND, K. H. 2011. Multivariate methods in metabolomics–from pre-processing to dimension reduction and statistical analysis. *TrAC Trends in Analytical Chemistry*, 30, 827-841.

LINDEQUE, J. Z., HIDALGO, J., LOUW, R. & VAN DER WESTHUIZEN, F. H. 2013. Systemic and organ specific metabolic variation in metallothionein knockout mice challenged with swimming exercise. *Metabolomics*, 9, 418-432.

LISEC, J., SCHAUER, N., KOPKA, J., WILLMITZER, L. & FERNIE, A. R. 2006. Gas chromatography mass spectrometry–based metabolite profiling in plants. *Nature protocols*, 1, 387-396.

LU, W., BENNETT, B. D. & RABINOWITZ, J. D. 2008. Analytical strategies for LC–MS-based targeted metabolomics. *Journal of Chromatography B*, 871, 236-242.

LUEDEMANN, A., STRASSBURG, K., ERBAN, A. & KOPKA, J. 2008. TagFinder for the quantitative analysis of gas chromatography—mass spectrometry (GC-MS)-based metabolite profiling experiments. *Bioinformatics*, 24, 732-737.

LUO, J., SCHUMACHER, M., SCHERER, A., SANOUDOU, D., MEGHERBI, D., DAVISON, T., SHI, T., TONG, W., SHI, L. & HONG, H. 2010. A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. *The pharmacogenomics journal*, 10, 278-291.

MADSEN, R., LUNDSTEDT, T. & TRYGG, J. 2010. Chemometrics in metabolomics—a review in human disease diagnosis. *Analytica Chimica Acta*, 659, 23-33.

MCINNES, J. 2013. Mitochondrial-associated metabolic disorders: foundations, pathologies and recent progress. *Nutrition & Metabolism*, 10, 63.

MONTEIRO, M., CARVALHO, M., BASTOS, M. & GUEDES DE PINHO, P. 2013. Metabolomics analysis for biomarker discovery: advances and challenges. *Current medicinal chemistry*, 20, 257-271.

MORGAN, P., HIGDON, R., KOLKER, N., BAUMAN, A., ILKAYEVA, O., NEWGARD, C., KOLKER, E., STEELE, L. & SEDENSKY, M. 2015. Comparison of proteomic and metabolomic profiles of mutants of the mitochondrial respiratory chain in Caenorhabditis elegans. *Mitochondrion*, 20, 95-102.

MUNNICH, A., RÖTIG A, CORMIER-DAIRE V & RUSTIN P 2013. Clinical Presentation of Respiratory Chain Deficiency. *The Online Metabolic and Molecular Bases of Inherited Disease*, 10.

NAIDU, K. A. 2003. Vitamin C in human health and disease is still a mystery? An overview. *Nutrition Journal*, 2, 7.

NIKOLIC, S. B., SHARMAN, J. E., ADAMS, M. J. & EDWARDS, L. M. 2014. Metabolomics in hypertension. *Journal of hypertension*, 32, 1159-1169.

PARIDA, S. K. & KAUFMANN, S. H. 2010. The quest for biomarkers in tuberculosis. *Drug discovery today*, 15, 148-157.

PFEFFER, G., MAJAMAA, K., TURNBULL, D. M., THORBURN, D. & CHINNERY, P. F. 2012. Treatment for mitochondrial disorders. *Cochrane Database Syst Rev*, 4.

PIECZENIK, S. R. & NEUSTADT, J. 2007. Mitochondrial dysfunction and molecular pathways of disease. *Experimental and molecular pathology*, 83, 84-92.

REINECKE, C. J., KOEKEMOER, G., VAN DER WESTHUIZEN, F. H., LOUW, R., LINDEQUE, J. Z., MIENIE, L. J. & SMUTS, I. 2012. Metabolomics of urinary organic acids in respiratory chain deficiencies in children. *Metabolomics*, 8, 264-283.

RHEE, E. P. & GERSZTEN, R. E. 2012. Metabolomics and cardiovascular biomarker discovery. *Clinical chemistry*, 58, 139-147.

RODENBURG, R. J. 2011. Biochemical diagnosis of mitochondrial disorders. *Journal of inherited metabolic disease*, 34, 283-292.

RÖTIG, A. & MUNNICH, A. 2003. Genetic features of mitochondrial respiratory chain disorders. *Journal of the American Society of Nephrology*, 14, 2995-3007.

SACCENTI, E., HOEFSLOOT, H. C., SMILDE, A. K., WESTERHUIS, J. A. & HENDRIKS, M. M. 2014. Reflections on univariate and multivariate analysis of metabolomics data. *Metabolomics*, 10, 361-374.

SAEED, S. & SINGER, M. 2013. Mitochondria—key roles in sepsis. *Réanimation*, 22, 352-358.

SARASTE, M. 1999. Oxidative phosphorylation at the fin de siecle. *Science*, 283, 1488-1493.

SCHAEFER, A. M., MCFARLAND, R., BLAKELY, E. L., HE, L., WHITTAKER, R. G., TAYLOR, R. W., CHINNERY, P. F. & TURNBULL, D. M. 2008. Prevalence of mitochondrial DNA disease in adults. *Annals of neurology*, 63, 35-39.

SCHAEFER, A. M., TAYLOR, R. W., TURNBULL, D. M. & CHINNERY, P. F. 2004. The epidemiology of mitochondrial disorders—past, present and future. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, 1659, 115-120.

SCHAPIRA, A. H. 2006. Mitochondrial disease. *The Lancet*, 368, 70-82.

SCHULTZ, A. W., WANG, J., ZHU, Z.-J., JOHNSON, C. H., PATTI, G. J. & SIUZDAK, G. 2013. Liquid Chromatography Quadrupole Time-of-Flight Characterization of Metabolites Guided by the METLIN Database. *Nature protocols*, 8, 451.

SCHYMANSKI, E. L., JEON, J., GULDE, R., FENNER, K., RUFF, M., SINGER, H. P. & HOLLENDER, J. 2014. Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Environmental science & technology*, 48, 2097-2098.

SHAH, V., CASTRO-PEREZ, J. M., MCLAREN, D. G., HERATH, K. B., PREVIS, S. F. & RODDY, T. P. 2013. Enhanced data-independent analysis of lipids using ion mobility-TOFMSE to unravel quantitative and qualitative information in human plasma. *Rapid Communications in Mass Spectrometry*, 27, 2195-2200.

SHAHAM, O., SLATE, N. G., GOLDBERGER, O., XU, Q., RAMANATHAN, A., SOUZA, A. L., CLISH, C. B., SIMS, K. B. & MOOTHA, V. K. 2010. A plasma signature of human mitochondrial disease revealed through metabolic profiling of spent media from cultured muscle cells. *Proceedings of the National Academy of Sciences*, 107, 1571-1575.

SHAHAM, O., WEI, R., WANG, T. J., RICCIARDI, C., LEWIS, G. D., VASAN, R. S., CARR, S. A., THADHANI, R., GERSZTEN, R. E. & MOOTHA, V. K. 2008. Metabolic profiling of the human response to a glucose challenge reveals distinct axes of insulin sensitivity. *Molecular Systems Biology*, 4, 214.

SIM, K. G., CARPENTER, K., HAMMOND, J., CHRISTODOULOU, J. & WILCKEN, B. 2002. Acylcarnitine profiles in fibroblasts from patients with respiratory chain defects can resemble those from patients with mitochondrial fatty acid [beta]-oxidation disorders. *Metabolism*, 51, 366-371.

SKLADAL, D., HALLIDAY, J. & THORBURN, D. R. 2003. Minimum birth prevalence of mitochondrial respiratory chain disorders in children. *Brain*, 126, 1905-1912.

SMITH, C. A., O'MAILLE, G., WANT, E. J., QIN, C., TRAUGER, S. A., BRANDON, T. R., CUSTODIO, D. E., ABAGYAN, R. & SIUZDAK, G. 2005. METLIN: a metabolite mass spectral database. *Therapeutic drug monitoring*, 27, 747-751.

SMOLINSKA, A., BLANCHET, L., BUYDENS, L. M. & WIJMENGA, S. S. 2012. NMR and pattern recognition methods in metabolomics: from data acquisition to biomarker discovery: a review. *Analytica chimica acta*, 750, 82-97.

SMUTS, I., LOUW, R., DU TOIT, H., KLOPPER, B., MIENIE, L. J. & VAN DER WESTHUIZEN, F. H. 2010. An overview of a cohort of South African patients with mitochondrial disorders. *Journal of inherited metabolic disease*, 33, 95-104.

SMUTS, I. & VAN DER WESTHUIZEN, F. H. 2010. Mitochondrial disorders - diagnostic approached and their application in the South African context. *South African Paediatric Review*, 7, 6-15

SMUTS, I., VAN DER WESTHUIZEN, F. H., LOUW, R., MIENIE, L. J., ENGELKE, U. F., WEVERS, R. A., MASON, S., KOEKEMOER, G. & REINECKE, C. J. 2013. Disclosure of a putative biosignature for respiratory chain disorders through a metabolomics approach. *Metabolomics*, 9, 379-391.

SPINAZZOLA, A. & ZEVIANI, M. 2009. Disorders from perturbations of nuclear-mitochondrial intergenomic cross-talk. *Journal of internal medicine*, 265, 174-192.

STYCZYNSKI, M. P., MOXLEY, J. F., TONG, L. V., WALTHER, J. L., JENSEN, K. L. & STEPHANOPOULOS, G. N. 2007. Systematic identification of conserved metabolites in GC/MS data for metabolomics and biomarker discovery. *Analytical Chemistry*, 79, 966-973.

SUE, C.M. & SCHON, E.A. 2000. Mitochondrial respiratory chain diseases and mutations in nuclear DNA: A promising start? *Brain pathology*, 10(3):442-450.

SUMNER, L. W., AMBERG, A., BARRETT, D., BEALE, M. H., BEGER, R., DAYKIN, C. A., FAN, T. W.-M., FIEHN, O., GOODACRE, R. & GRIFFIN, J. L. 2007. Proposed minimum reporting standards for chemical analysis. *Metabolomics*, 3, 211-221.

SUOMALAINEN, A. 2011a. Biomarkers for mitochondrial respiratory chain disorders. *Journal of inherited metabolic disease*, 34, 277-282.

SUOMALAINEN, A. Therapy for mitochondrial disorders: little proof, high research activity, some promise. *Seminars in Fetal and Neonatal Medicine*, 2011b. Elsevier, 236-240.

T'KINDT, R., MORREEL, K., DEFORCE, D., BOERJAN, W. & VAN BOCXLAER, J. 2009. Joint GC–MS and LC–MS platforms for comprehensive plant metabolomics: Repeatability and sample pre-treatment. *Journal of Chromatography* B, 877, 3572-3580.

TAUTENHAHN, R., BÖTTCHER, C. & NEUMANN, S. 2008. Highly sensitive feature detection for high resolution LC/MS. *BMC bioinformatics*, 9, 504.

TAVAZZI, B., LAZZARINO, G., LEONE, P., AMORINI, A. M., BELLIA, F., JANSON, C. G., DI PIETRO, V., CECCARELLI, L., DONZELLI, S. & FRANCIS, J. S. 2005. Simultaneous high performance liquid chromatographic separation of purines, pyrimidines, N-acetylated amino acids, and dicarboxylic acids for the chemical diagnosis of inborn errors of metabolism. *Clinical biochemistry*, 38, 997-1008.

TAYLOR, R. W. & TURNBULL, D. M. 2005. Mitochondrial DNA mutations in human disease. *Nature Reviews Genetics*, 6, 389-402.

VAN DEN BERG, R. A., HOEFSLOOT, H. C., WESTERHUIS, J. A., SMILDE, A. K. & VAN DER WERF, M. J. 2006. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC genomics*, 7, 142.

VENTER, L., LINDEQUE, Z., VAN RENSBURG, P. J., VAN DER WESTHUIZEN, F., SMUTS, I. & LOUW, R. 2015. Untargeted urine metabolomics reveals a biosignature for muscle respiratory chain deficiencies. *Metabolomics*, 11, 111-121.

VERGANO, S. S., RAO, M., MCCORMACK, S., OSTROVSKY, J., CLARKE, C., PRESTON, J., BENNETT, M. J., YUDKOFF, M., XIAO, R. & FALK, M. J. 2014. In vivo metabolic flux profiling with stable isotopes discriminates sites and quantifies effects of mitochondrial dysfunction in C. elegans. *Molecular genetics and metabolism*, 111, 331-341.

VINAIXA, M., SAMINO, S., SAEZ, I., DURAN, J., GUINOVART, J. J. & YANES, O. 2012. A guideline to univariate statistical analysis for LC/MS-based untargeted metabolomics-derived data. *Metabolites*, 2, 775-795.

VO, T. D., LEE, W. P. & PALSSON, B. O. 2007. Systems analysis of energy metabolism elucidates the affected respiratory chain complex in Leigh's syndrome. *Molecular genetics and metabolism*, 91, 15-22.

WALLACE, D. C. & FAN, W. 2010. Energetics, epigenetics, mitochondrial genetics. *Mitochondrion*, 10, 12-31.

WISHART, D. S., TZUR, D., KNOX, C., EISNER, R., GUO, A. C., YOUNG, N., CHENG, D., JEWELL, K., ARNDT, D. & SAWHNEY, S. 2007. HMDB: the human metabolome database. *Nucleic acids research*, 35, D521-D526.

WOLF, N. I. & SMEITINK, J. A. 2002. Mitochondrial disorders A proposal for consensus diagnostic criteria in infants and children. *Neurology*, 59, 1402-1405.

XIA, J., BROADHURST, D. I., WILSON, M. & WISHART, D. S. 2013. Translational biomarker discovery in clinical metabolomics: an introductory tutorial. *Metabolomics*, 9, 280-299.

XIA, J., PSYCHOGIOS, N., YOUNG, N. & WISHART, D. S. 2009. MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic acids research*, 37, W652-W660.

XU, F., ZOU, L. & ONG, C. N. 2010. Experiment-originated variations, and multi-peak and multi-origination phenomena in derivatization-based GC-MS metabolomics. *TrAC Trends in Analytical Chemistry*, 29, 269-280.

YANES, O., TAUTENHAHN, R., PATTI, G. J. & SIUZDAK, G. 2011. Expanding coverage of the metabolome for global metabolite profiling. *Analytical chemistry*, 83, 2152-2161.