The metabolic profile of clinical and immunogenetic factors linked to HIV progression

E Jansen van Rensburg

orcid.org/0000-0001-8245-0432

Dissertation accepted in partial fulfilment of the requirements for the degree *Master of Science in Biochemistry* at the North-West University

Supervisor:	Dr AA Williams
Co-supervisor:	Prof DT Loots
Assistant Supervisor:	Prof T Ndung'u

Graduation May 2020

23498536

ACKNOWLEDGEMENTS

This dissertation is the result of three years' worth of hard work and overcoming challenges presented along the way. This study would not have been possible without our God who intricately created us and within us the ability to fight against evil. I thank my God for the inspiration, motivation, physical and financial support granted to me during the last three years.

My wife Euné, although you didn't understand a fraction of this study, you stood by my side, motivating, encouraging and supporting me. I appreciate your grace and patience with the late nights and mixed emotions. Finally, thank you for nine months of preparing the best graduation gift a man can ask for, our son Luke.

To my parents and sisters, thank you for all your support, advice and help. I would not have been able to keep a sane mind without all those G&T and braai weekends in Parys.

Dr Aurelia Williams, thank you for your guidance and support in these last three years. Although you were newly appointed at the NWU, you accepted the challenge to supervise this study. I have learnt from you in more ways than I could imagine. From your guidance, I've gained experience in practical laboratory aspects, writing aspects, interpersonal relations, positive thinking and last but not least, karaoke and paintball.

Prof Du Toit Loots, thank you for assisting me in this project, helping me figure out the things I would not have been able to on my own. Thank you for the social support, building chats and drinking coffee during breaks.

Prof Thumbi Ndung'u, thank you for your help in acquiring samples for this project and your input on the *HLA-B* alleles.

I thank all my fellow students and co-workers for always having time for coffee, tea or a beer. I especially want to acknowledge my tea partners: Zinandre Stander, Monique Combrinck (soon to be Opperman), Tiaan van Zyl, Anandi Rautenbach and Varushka Acton.

Finally, I want to thank the National Research Foundation (NRF) for the funds provided toward this project. (The financial assistance of the NRF towards this research is hereby acknowledged. The opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF).

ABSTRACT

HIV disease progression is generally defined by the time it takes an individual to progress from primary HIV infection to the acquired immunodeficiency syndrome (AIDS). CD4 T-cell count and plasma viral load are validated clinical indicators of disease progression. These parameters are, however, not reliable, varying significantly across HIV-infected persons. The metabolic and biological reasons underlying the variation in these markers of disease progression are not entirely known, but immunogenetic factors are known to contribute significantly.

This study compared the plasma metabolic profile (n=96) of untreated HIV positive participants (n=53) presenting clinical and immunogenetic factors previously linked to HIV disease progression. Samples were extracted, derivatised and analysed on the Leco Pegasus 4D system. The samples of participants with high CD4 (500-800 cells/µl) and low CD4 (<250 cells/µl) counts were compared. The samples of participants with median CD4 counts (350-499 cells/µl) with a non-significant versus significant negative correlation with time (termed non-progressors and progressors) and median CD4 counts (350-499 cells/µl) with protective vs non-protective *HLA-B* alleles, respectively were compared.

The samples of participants with **low CD4 counts** had decreased amino acids, fatty acids and carbohydrates indicating increased protein catabolism and a reduction in the intake and absorption of branched-chain amino acids (BCAAs). Decreased levels of uridine and an increase in microbial metabolites suggests continued viral replication and microbial dysbiosis. The samples of participants with **significant negative correlation between CD4 count and time** presented with less metabolic variation implying CD4 count over time to not significantly impact on the host metabolism. The samples of participants with **non-protective *HLA-B* alleles** reflected a general increase in amino acids, fatty acids, carbohydrates and microbial metabolites.

The clinical factor, CD4 was associated with distinct metabolic changes compared to the change in CD4 over time, with trends suggestive of a shift towards the use of these metabolites for energy metabolism. The samples of participants with non-protective *HLA-B* alleles revealed metabolic changes indicative of immune activation and microbial dysbiosis. Although samples stratified according to clinical and immunogenetic factors displayed distinct metabolite profiles implying varied mechanisms to contribute to differential HIV disease progression, groups with a "poorer" outcome generally showed features with some similarity. While clinical, immune, genetic and other factors have been used to define patient prognosis, a more holistic view into differential disease progression in these patients may benefit from the inclusion of a metabolic component.

Keywords: HIV/AIDS; progression; metabolomics; CD4; viral load; clinical; immunogenetic; *HLA-B*

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	Meaning
1H NMR	Proton Nuclear Magnetic Resonance
2D-GC	Two-Dimensional Gas Chromatography
3-PHB	3-Phenylbutyric acid
AA	Amino acids
AFROX	African Oxygen Limited
AIDS	Acquired Immunodeficiency Syndrome
AMP	Adenosine monophosphate
AMPK	Adenosine monophosphate kinase
ART	Antiretroviral Therapy
ATP	Adenosine triphosphate
ATR-FTIR	Attenuated Total Reflectance Fourier-Transform
AZT	Azidothymidine
BCAA	Branched-chain amino acid
BCAT	Branched-chain aminotransferase
BREC	Biomedical research ethics committee
BSTFA	N, O-Bis(trimethylsilyl)trifluoroacetamide
CA	Carbohydrates
cART	Combination Antiretroviral Therapy
CAS	Chemical Abstracts Service
CCR5	Cysteine-cysteine chemokine receptor 5
CD4	Cluster of Differentiation 4
CD8	Cluster of Differentiation 8
CV	Coefficient of Variance
DNA	Deoxyribonucleic acid
DP	Data Processing
EC	Elite controllers
EI	Electron Impact Ionisation
EI-MS	Electron Impact ionisation Mass Spectrometry
ELISA	Enzyme-Linked Immunosorbent Assay
ES	Effect size
ESI	Electrospray Ionization
FA	Fatty acids
FDA	Food and drug administration
GALT	Gut-Associated Lymphoid Tissue
GC	Gas Chromatography
GC-MS	Gas Chromatography-Mass Spectrometry
GCxGC	Two-Dimensional Gas Chromatography
GCxGC-TOFMS	Two-Dimensional Gas Chromatography Time Of Flight Mass Spectrometry
HAART	Highly Active Antiretroviral Therapy
HI	Human immunodeficiency
HIV	Human Immunodeficiency Virus
HLA	Human Leucocyte Antigen

Abbreviation	Meaning
HMDB	Human Metabolome Database
IDO	Indoleamine 2,3-dioxygenase
IL	Interleukin
IMGT	International immunogenetics information system
IS	Internal Standard
IUPAC	International Union of Pure and Applied Chemistry
KIR	Killer-cell immunoglobulin-like receptor
LC	Liquid Chromatography
LC-MS	Liquid Chromatography-Mass Spectrometry
LDL	Low-density lipoprotein
LOOCV	Leave one out cross-validation
LTNP	Long-term non-progressors
LTSP	Long-term slow-progressors
MADD	Multiple Acyl-coenzyme A Dehydrogenase Deficiency
MHC	Major Histocompatibility Complex
MMCA	Metabolite-metabolite correlation analysis
MS	Mass Spectrometry
MS^E	Mass Spectrometry at different Fragmentation energies
MS^n	Mass Spectrometry tree of fragments
MSTUS	Mass Spectrum Total Useful Signal
MW	Mann-Whitney
NIST	National Institute of Standards and Technology
NMR	Nuclear Magnetic Resonance
NNRTI	Non-nucleoside reverse transcriptase inhibitors
NP	Non-progressors
NRTI	Nucleoside Reverse Transcriptase Inhibitor
NWU	North-West University
PCA	Principal component analysis
PCR	Polymerase Chain Reaction
PI	Protease inhibitors
PID	Patient/Participant Identifier
PLS-DA	Partial least squares discriminant analysis
PMTCT	Prevention of Mother-To-Child Transmission
QC	Quality control
QC-CV	Quality control Coefficient of variance
RNA	Ribonucleic Acid
ROS	Reactive oxygen species
RP	Rapid progressors
RSD	Relative standard deviation
RT	Retention time
SAB	South African Black
SAC	South African Caucasian
SAI	South African Indian
SAM	South African Mixed Ancestry
SD	Standard deviation

Abbreviation	Meaning
SIV	Simian Immunodeficiency Virus
SOP	Standard Operating Procedure
SSO	Sequence-specific oligonucleotide
TIC	Total ion chromatogram
TLR	Toll-Like Receptor
TMCS	Trimethylchlorosilane
TMS	Trimethylsilane
TOF	Time of Flight
TOFMS	Time Of Flight Mass Spectrometry
tRNA	Transfer Ribonucleic Acid
U	Unknown
UKZN	The University of KwaZulu-Natal
UNAIDS	The Joint United Nations Programme on HIV/AIDS
UTT	Universal test and treat
VIP	Variable importance in projection
VL	Viral Load
WHO	World Health Organisation

INTRODUCTION

1.1 Background and motivation

During 2015, the Joint United Nations Programme on HIV/AIDS launched the 90-90-90 target for 2020 aimed at ending the AIDS pandemic. Although significant progress has been made towards achieving these targets, only two-thirds of people living with HIV know their status, 77% have access to antiretroviral therapy, and 82% of people taking treatment have suppressed viral loads (HIV/AIDS, 2017). HIV/AIDS statistics from 2018 further highlights the slow response to the pandemic, with 1.7 million people being newly infected with HIV, increasing the global population of people living with the virus to 37.9 million. During this time, South Africa accounted for nearly 20% (7.2 million) of all global infections, which claimed 940 000 lives. These statistics highlight just how serious of a global health concern HIV/AIDS still is, necessitating studies which better characterise the disease.

HIV is transmitted mainly through sexual intercourse. The HI-virus upon infecting its host results in a cascade of inflammatory and cell-mediated immune responses which subsequently impacts on the host metabolism. HIV infection is characterised into three stages, determined by viral load (VL) and cluster of differentiation 4 (CD4) count, i.e. (1) primary infection where the virus rapidly multiplies, and CD4 count rapidly decreases, (2) the asymptomatic /chronic stage where viral load stabilises and CD4 count steadily decreases, and (3) the AIDS stage where viral load increases rapidly and CD4 counts drop below 200 cells/µl blood. The typical clinical staging is not representative of all cases as individuals are heterogeneous in their response to HIV exposure and infection, complicating the management of the disease. While the CD4 cell count and VL are used in the clinical setting to monitor patient prognosis, these parameters are prone to variation/error. There is thus a need for complementary markers with which to assess patient well-being and prognosis.

To alleviate the burden of HIV infection, antiretroviral therapy (ART) or combinations thereof referred to as highly active antiretroviral therapy (HAART) is used to maintain these HIV-infected individuals in the asymptomatic phase of the disease. ART/HAART represents an artificial, non-natural way of slowing HIV disease progression. However, cases representing natural biological control of the disease, and displaying slow progression phenotypes have been observed. Many factors which impact the natural pathology of HIV and slow disease progression have been described. In this regard, the immune and genetic parameters are mainly reported.

To date, several definitions exist for those individuals capable of maintaining a moderate to high CD4 count in the absence of treatment. The patient's ability to control viral load and CD4 count is used to classify them as controllers and/or slow/non-progressors. These definitions share a protective phenotype but differ in the timeframe used to define each, i.e. some definitions are defined based on the ability of the patient to present the protective phenotype over five years and in other instances over 10-plus years. While many studies have been done to uncover the mechanisms explaining the heterogeneity in HIV disease progression, the cohorts have mainly been defined using clinical and/or immunogenetic data (eg. CD4, VL, cysteine-cysteine chemokine receptor 5 [*CCR5*], human leukocyte antigen [*HLA*] and killer-cell immunoglobulin-like receptor [*KIR*] genotype, etc). The use of metabolomics (i.e. the study of small organic molecules that form part of the chemical reactions in living organisms) to uncover mechanisms associated with HIV disease progression is however limited.

Gupta and Gupta (2004) describe several factors affecting HIV disease progression. These include but is not limited to viral strain, subtype, host immune status and environmental factors. Identifying physiological or biochemical changes that can differentiate progressors and non-progressors at baseline, would undoubtedly aid in characterising prognosis while the individual is still in a relatively healthy state. Madec *et al.* (2009) established long term slow progressor (LTSP) status of their cohort within a year of HIV infection using CD4. This early identification using clinical measurements of progression serves as a basis to measure early metabolic changes that are predictive of HIV disease progression. The optimisation of disease management strategies in a South-African context will additionally benefit from understanding the physiological and biochemical mechanisms affecting HIV disease progression. This gap can potentially be addressed through an untargeted metabolomics approach which measures the metabolic responses of living systems to biological stimuli (Schoeman & Loots, 2011).

Because metabolites are the products of the transcriptome and the proteome, metabolomics may be a better way to study the metabolic reprogramming caused by HIV-induced immune activation. An untargeted metabolomics approach to study HIV disease progression may reveal the downstream effects of the immunogenes that predict HIV disease progression. Two-dimensional gas chromatography coupled to time of flight mass spectrometry (GCxGC-TOFMS) is an especially well-suited analytical platform for analysing chemically complex samples with a high sample dimensionality (Schoeman & Loots, 2011).

Many metabolomics studies have been conducted to characterise the metabolic fingerprint of HIV, and its treatment. These studies mostly found metabolites linked to inflammation as characteristic of the disease. Additionally, several researchers reported metabolites of gut microbial origin and mitochondrial damage. Recent targeted metabolomics studies identified metabolites capable of distinguishing between samples representing different rates of progression (Scarpelini *et al.*, 2016; Zhang *et al.*, 2018). Limitations to these prior investigations includes the small numbers of metabolites (targeted metabolomics approach) and samples (n=10) that were analysed. To date, only one study employing an untargeted metabolomics approach has been applied to characterise HIV disease progression (Scarpelini et al., 2016). An untargeted metabolomics analysis of HIV progressors and non/slow-progressors with clearly defined criteria for progression based on clinical and immunogenetic markers will assist in gaining a better understanding of the underlying mechanisms of disease progression from a metabolic perspective.

A better understanding of the mechanisms of natural slow/non-progression is beneficial for improving the quality of life for those individuals living with a positive HIV diagnosis. With the knowledge gained, more accurate prognosis and disease monitoring strategies with earlier intervention to limit progression is expected to follow. Additionally, a better understanding of these mechanisms may inform on new treatment strategies focussed on modulating the immune system and/or metabolism to transform normal and rapid progressors into slow/non-progressors without the use of ART/HAART. The factors identified as protective to the slow-/non-progressors will also possibly contribute to improved vaccine design strategies.

This study will provide a greater understanding of HIV disease progression in South Africa, which is a crucial step towards attempting to reduce this pandemic. This study will add new knowledge about HIV-induced metabolic markers associated with HIV disease progression, which may better describe mechanisms of natural control of the virus. More specifically, we will look at the influence of clinical and host immunogenetic factors on the metabolism and how this impacts on HIV disease progression.

1.2 Aims and objectives

1.2.1 Aim

In this study, we aim to investigate the altered metabolic profiles associated with various clinical and immunogenetic factors linked to untreated HIV disease progression by applying an untargeted two-dimensional gas chromatography time-of-flight mass spectrometry (GCxGC-TOFMS) metabolomics approach, to better understand the metabolic mechanisms associated with differential disease progression.

The sub-studies of this aim is to compare the plasma metabolic profiles of untreated HIV positive participants with:

1. High CD4 (500-800 cells/µl) vs. low CD4(<250 cells/µl) counts;
2. Median CD4 counts (350-499 cells/µl) with a non-significant vs significantly negative correlation with time (termed non-progressors and progressors);
3. Median CD4 counts (350-499 cells/µl) with protective vs non-protective *HLA-B* alleles, respectively.

1.2.2 Objectives

The following objectives were formulated to achieve the aforementioned aim:

1. Find a collaborator willing to share samples of participants with longitudinal CD4 count records (in relation to sub-study 2), while these participants should have samples available at different CD4 counts, low (<250 cells/µl]) and high (500-800 cells/µl) for sub-study 1 and median [350-499 cells/µl] for sub-studies 2 and 3.
2. Select the most suitable samples from each participant for optimal investigation of the aim.
3. Prepare a quality control (QC) sample and standardise the analysis method.
4. Analyse the samples using the standardised method.
5. Process the raw data.
6. Group the samples according to the respective sub-studies and perform statistics on the groups to differentiate the metabolic profiles.
7. Interpret the metabolic profiles of the samples from the respective sub-studies and provide hypotheses on metabolic changes.

1.3 Scope and considerations of the project

The availability of specific samples was a major limitation to this project as well as the experimental design. The most significant role player in this regard was the universal test and treat (UTT) policy, which eliminates the opportunity to recruit new untreated participants showing differential disease progression. Treatment-naïve HIV-infected individuals are crucial for the investigation of the natural untreated metabolic profile. Sample collection, therefore, had to take place before 1 September 2016 (the date at which the implementation of UTT commenced) to be considered for this project. Ethical considerations, policies and laws have made it impossible to recruit HIV-infected participants and study their progression without treatment administration. Additionally, at least four years' worth of CD4 measurement records before this date had to accompany the samples for assessment of clinical progression.

The scope of the project was to investigate the metabolic profiles of samples from participants presenting different clinical and immunogenetic parameters conventionally used to measure HIV disease progression. To investigate the metabolic profile, as many metabolites as possible must be detected and quantified. No single analytical platform developed to date has the capacity to detect and quantify all metabolites in a biological sample. Multiple targeted analyses on different platforms provides a larger coverage of the metabolome, but will not be able to inform on those metabolites not targeted. For this reason, among others, it was decided to use an untargeted approach to analyse samples. The Leco Pegasus 4D GCxGC-TOFMS analytical system has great resolution, increased sensitivity and would provide comprehensive coverage of most metabolites. An extraction and analysis method for serum was previously optimised by Parihar *et al.* (2017). Due to the availability of this established method, it was beyond the scope of this project to design a new method, and only minor modifications were made to the split ratio and detector voltage (discussed in 4.4.1.) of the GC method.

1.4 Structure and outputs

This dissertation is written to comply with the requirements of the North-West University (NWU), Potchefstroom Campus, South Africa, for the completion of the degree Magister Scientae (Biochemistry) in dissertation format.

Chapter 1 is an **introduction** highlighting the critical aspects of this study and the gap it aims to address through the aims and objectives.

Chapter 2 is a **literature review** focussing on HIV infection, disease progression and metabolomics.

Chapter 3 describes the **experimental design** of this study, including a description of the participant and sample selection protocol.

Chapter 4 describes the reagents, **materials and methods** used during this study.

Chapter 5 discusses the **results** obtained, providing an overall interpretation to the meaning thereof.

Chapter 6 provides a **conclusion** considering all the data gathered in this study. **Future considerations** for HIV-based progression studies are also provided.

Findings from this study were presented at the launch of Metabolomics South Africa at the Innovation Hub in Pretoria as an oral presentation (title: “The metabolomics of treatment-naïve HIV-infected progressors and non-progressors”) as well as at the 9th SA AIDS conference in Durban as a poster (title: “Characterising HIV progression using metabolomics”). This study also provided the opportunity to attend an advanced metabolomics workshop in Pretoria and from experience gained through the workshop as well as this study, an opportunity arose for me to present and facilitate a series of wet-lab experiments at an introductory metabolomics workshop hosted by the NWU.

As a student of this project, I gained a lot of technical, administrative and interpersonal skills. Due to the nature of this study, I received training on the preparation and handling of biosamples under sterile conditions as well as the safe disposal thereof. I had the opportunity to work on one of the most advanced analytical platforms available and spent time with application specialists learning even more. I mastered various extraction techniques and was also introduced to flow cytometry. I had the opportunity to present data from another project that I was part of at the Australian & New Zealand metabolomics conference in Auckland in 2018 (title: “The GCxGC-TOF MS metabolic profile of HIV-infected sera and its association with markers of cardiovascular disease”).

1.5 Study contributions

The primary author of this dissertation is Emile Jansen van Rensburg. The contributions of the co-authors, co-workers and collaborators made towards this work are summarised below:

Name	Role	Contribution
Emile Jansen van Rensburg	Author	Responsible for the conceptualising, planning, execution and reporting of this study along with the study leader.
Dr Aurelia Williams	Co-author	Study leader: Conceptualised, coordinated and supervised all aspects of the study. She was also responsible for the administrative aspects of the study.
Prof Du Toit Loots	Co-author	Co-supervisor: Assisted with the design and planning of the study as well as data interpretation and final write-up.
Prof Thumbi Ndung'u	Co-author	Co-supervisor: Assisted with the design and planning of the study. Provided the samples and assisted with the sample selection. Proof read outputs which emanated from the study.
Mrs Derylize Beukes-Maasdorp	Co-worker	Provided training on sample handling, the extraction of samples and the analysis of these on the GCxGC-TOFMS system.
Dr Mari van Reenen	Co-worker	Assisted with the design of the study as well as statistical analysis and interpretation of the data.
The University of KwaZulu-Natal	Collaborator	Provided the plasma samples used in this metabolomics study

LITERATURE REVIEW

This dissertation focusses on the altered metabolic profile of plasma collected from an HIV positive study population presenting different factors linked to HIV disease progression. In light of this, the following literature review aims to present an overview of various biological aspects of HIV infection, the clinical stages of HIV infection, HIV disease progression, as well as the metabolomics of HIV.

2.1 HIV virology and pathology

Acquired immunodeficiency syndrome (AIDS) was first reported in humans in 1981, followed by the identification of its causative agent, the human immunodeficiency virus (HIV), in 1983 (Barré-Sinoussi *et al.*, 1983; Gottlieb *et al.*, 1981). HIV is believed to have been transmitted to humans decades ago when nonhuman primates carrying the simian immunodeficiency virus (SIV) was hunted as a source of food. Although the interspecies transmission of this virus is not plausible, multiple virus species, suitable for possible human infection, have evolved and been transmitted across different species. HIV and SIV, both have high mutation and recombination rates, which together with multiple zoonotic transmissions, allowed it to evolve into a diverse multi-strain virus with massive genetic heterogeneity and variability (Gao *et al.*, 1999; Hemelaar, 2012; Sakuma & Takeuchi, 2012).

Since its discovery, the pandemic has spread worldwide, claiming more than 32 million lives to date. At the end of 2018, approximately 37.9 million people globally were HIV positive, with 1.7 million new infections in 2018. The World Health Organization (2019) estimates that only 79% of infected people know their HIV status. Therefore, the estimated number of infections could be even higher. These statistics highlight that HIV/AIDS remains a global health priority.

HIV reverse transcriptase has a high error rate, which leads to replication errors and subsequently, huge genetic diversity. Many viral isolates have been phylogenetically analysed which lead to the classification of HIV into species, types, groups, subtypes, sub-sub types, circulating recombinant forms and unique recombinants. Two HIV species exist, HIV-1 and HIV-2, of which HIV-1 is more virulent and prevalent globally and thus much better characterised than HIV-2. HIV-1 will be investigated in this study; therefore, HIV-2 strains will not be discussed. HIV-1 strains are classified into three groups: M, N and O. Strain M is the most prevalent worldwide and further divided into subtypes. Nine subtypes of HIV-1 strain M exists, of which subtypes A and C are most common in Africa (Peeters *et al.*, 2003). For this study, all further reference to HIV will be with respect to HIV-1 strain M Subtype C.

Although minor differences exist between strains and subtypes, the general structure and replication cycle of HIV remains the same. HIV is an enveloped single-stranded ribonucleic acid (RNA) virus of the group lentiviruses (Gao *et al.*, 1999). Like other lentiviruses, HIV infects and replicates inside its hosts, mainly infecting cells presenting the cluster of differentiation 4 (CD4) antigen, designated CD4+ cells.

HIV transmission mainly occurs through the exchange of bodily fluids which contain mature virus. Transmission via penetration through several mucosal surfaces, to eventually reach the CD4+ cells, accounts for most HIV infections, with more than 80% of infections occurring in this way. Transmission can occur through either cell-free virus or cell-associated virus migrating through the mucosal membranes to the target cells (Barreto-de-Souza *et al.*, 2014; Cohen *et al.*, 2011). Once HIV reaches a CD4+ cell, the virus adheres to the cell.

Gp120 proteins on the viral envelope bind to CD4 receptors, linking the virus to the host cell membrane, which causes a conformational change in the viral surface protein Env that Gp120 forms part of. Cysteine-cysteine chemokine receptor 5 (*CCR5*) is used as a coreceptor to the conformed Env protein and triggers membrane fusion (Wilen *et al.*, 2012). Upon the release of the viral contents into the cytoplasm, reverse transcription begins. The viral RNA template, a host deoxyribonucleic acid (DNA) polymerase and a primer from the host's transporter RNAs (tRNAs) assemble and reverse transcription is initiated. RNase H enzymatically degrades the RNA in the RNA-DNA duplex except for a purine-rich sequence which serves as the primer for the reverse strand synthesis (Hu & Hughes, 2012). Integration of the complementary proviral DNA (cDNA) into human DNA is the next step in the HIV life cycle. Integration requires specific sequences on the ends of the cDNA to bind to the viral integrase and other proteins to form the pre-integration complex. Viral cDNA is then integrated into the host DNA at active gene sites and regional hotspots (Schröder *et al.*, 2002). Once integrated into the host genome, host transcription and translation produces all the components needed to create a new virus particle. The HIV-1 Gag polyprotein is responsible for virion assembly. Two copies of viral RNA, cellular tRNA, Env proteins, the viral enzymes and the Gag polyprotein assemble at the cell membrane. The assembly of components buds off the cell and produces an immature virus. Finally, the Gag polyprotein is cleaved into the viral capsid proteins which condense to stabilise the dimeric RNA genome. Env proteins migrate to the viral membrane, forming a mature virus which can infect neighbouring/nearby CD4+ cells (Sundquist & Kräusslich, 2012).

Inside the cells, viral replication occurs and more virus buds off into the extracellular space and quickly spreads to the draining lymphoreticular tissues (Cohen *et al.*, 2011) infecting any CD4 expressing cells along the way. For the first 7 to 21 days after infection, the viral RNA concentration remains too low to detect with qualitative polymerase chain reaction (PCR), but the viral concentration grows exponentially as new CD4+ cells are infected. In the 21 – 28 days following infection virus replication increases exponentially to more than a million copies of viral RNA per millilitre blood (McMichael *et al.*, 2009).

Only 2% of all lymphocytes are in circulation at any time. The remainder of the lymphocytes resides throughout the body in lymphoid organs such as the spleen, lymph nodes and gut-associated lymphoid tissue (GALT) (Blankson *et al.*, 2002). HIV virions spread through the circulatory system to these clusters of lymphocytes. Many of the lymphocytes express the CD4 and *CCR5* receptors needed for viral infection and are thus infected. Fortunately, the immune system is capable of launching a counter-attack on the virus by activating both innate and adaptive immune systems in response to infection. The innate immune system is likely a driving factor for immune activation through its activation by Toll-like receptors (TLRs). CD4+ T cells are activated by recognition of an antigen and undergo rapid clonal proliferation into effector CD4+ T cells. These cells are rapidly infected and have a high mortality rate. Infected cells have been shown to have a short half-life of less than one day as they die from immune responses or viral cytopathic effects (Blankson *et al.*, 2002; Chun & Fauci, 1999). A small number of these effector cells which have been infected by HIV manage to survive and differentiate into quiescent memory CD4+ T cells (Gasper *et al.*, 2014). Although initially believed to reside in the lymph nodes, new evidence suggests that these “genetic reservoir” memory T cells can occur in a variety of tissues including the lymphoid organs, central nervous system and the genitourinary tract (Blankson *et al.*, 2002). The cytotoxic T-cell response is both beneficial and detrimental as it may suppress viral replication, but fuel chronic T-cell activation.

The receptors on specific B-cells of the humoral immune system binds to the gp120 protein of HIV. Once bound, the B-cell engulfs the virus via endocytosis and digests it. Human leukocyte antigen class II (HLA-II) proteins then bind to various digested viral proteins and present them on the cell membrane. The receptors of T helper cells along with CD4 recognise some of the viral peptides presented by the B-cells. The activation thereof stimulates the secretion of B-cell activating cytokines such as Interleukin-2 (IL-2), IL-4 and IL-5. The activated B-cells proliferate into plasma cells which then secrete antibodies against the gp120 viral peptides.

A balance between viral turnover and the elicitation of an immune response characterises the viral set-point (McMichael *et al.*, 2009). Viral set-point establishes within three months of infection. Although a reduced viral load characterises the initial viral set-point, the actual viral set-point varies up to 1000 fold (1000 to 1 million copies of viral RNA per millilitre blood) between infected individuals (Fraser *et al.*, 2007). The reduced viral load alleviates the viral pressure on CD4+ cells while continuous immune activation via TLRs stimulates CD4+ proliferation. Subsequently, the decline in CD4+ cells slows down and stabilises.

CD4 cell counts in uninfected persons range from 330 cells/µl to 1610 cells/µl (Pantaleo & Fauci, 1996). Figure 2-1 shows a generalised progression curve for HIV-infected individuals defined by CD4 count and VL. Briefly, during acute infection, the virus rapidly infects new cells and multiplies, leading to a sharp increase in VL and decrease in CD4 cell count. During the clinical latency phase the immune response suppresses HIV replication leading to a decrease in VL and a slight increase in CD4+ T cells due to the alleviated pressure by HIV. This phase is generally the most prolonged phase of HIV infection (Siliciano & Greene, 2011). In subsequent years, HIV slowly spreads and decreases the remaining CD4 cell count. A diagnosis confirming AIDS occurs once an individual's CD4 count falls below 200 cells/µl.

Clinical latency in individuals varies, lasting as little as three years in some individuals and up to 10 years in others (Kumar, 2013). Long-lasting clinical latency consequently defined a population of HIV-infected individuals able to naturally control viral replication and maintain normal CD4 counts without intervention (Lu *et al.*, 2016). Due to the heterogeneous response to HIV exposure and infection, the typical clinical course outlined above does not hold true for all individuals. Those individuals who maintain moderate to high CD4 counts and low VL in the absence of treatment have been named as long term non-progressors (LTNPs) (Madec *et al.*, 2009) or long term slow-progressors (LTSP) (labelled late progressors in Figure 2-2) while others have decreased CD4 counts and increased VL and are known as rapid progressors (RP) (Mlisana *et al.*, 2014). Controllers, according to Grabar *et al.* (2017), are a subgroup of LTNPs with an undetectable viral load, while Rappocciolo *et al.* (2014) define controllers as having 50-2000 copies of viral RNA per ml blood, and elite controllers as having undetectable VL. Mandalia *et al.* (2011) define LTNPs as having low VL and controllers as having undetectable VL. There is no consensus regarding the definitions. Therefore, we defined our cohort based on the available clinical and immunogenetic data, in line with the definitions of previous literature. Controllers and LTNPs might have different mechanisms of controlling HIV, but some of these mechanisms may overlap since both groups yield favourable patient outcomes. Similarly, non-controllers and progressors may have shared mechanisms.

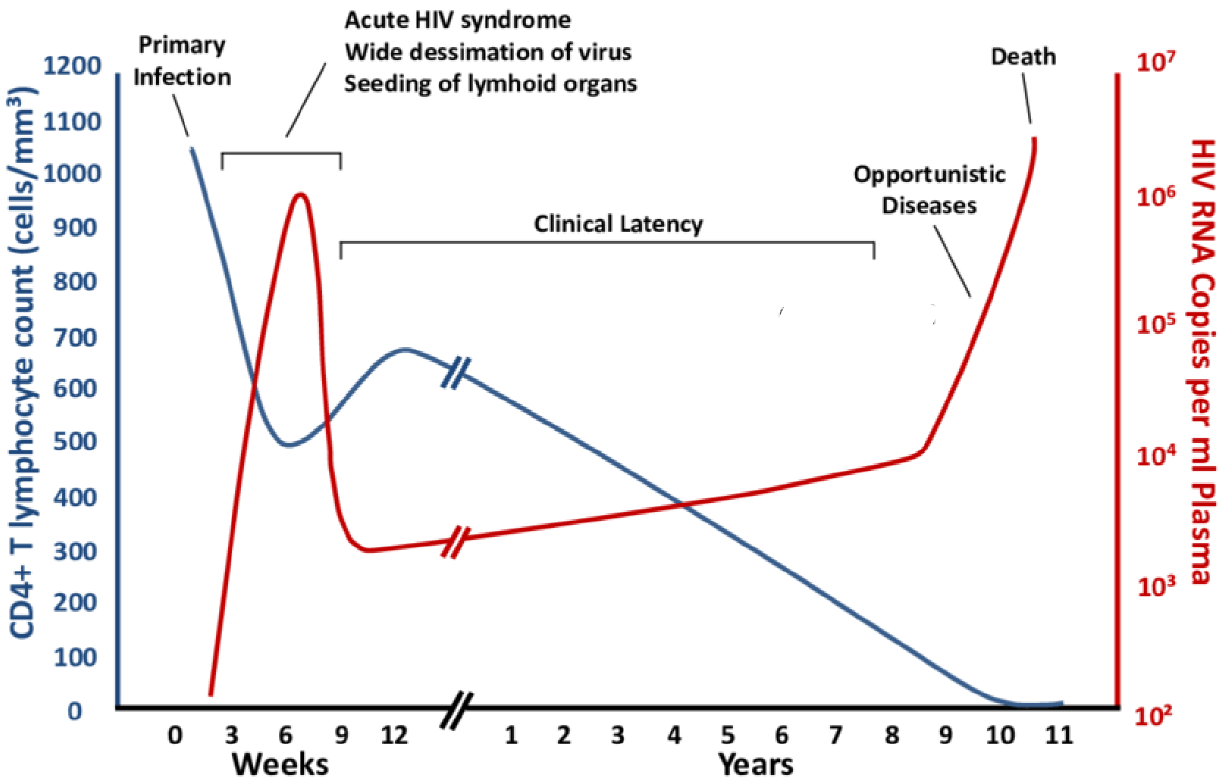

Figure 2-1: **The clinical course of HIV infection defined by CD4 count and viral load. Figure reproduced from Goovaerts (2015). (Used under fair dealing rights as described in the SA Copyright act)**

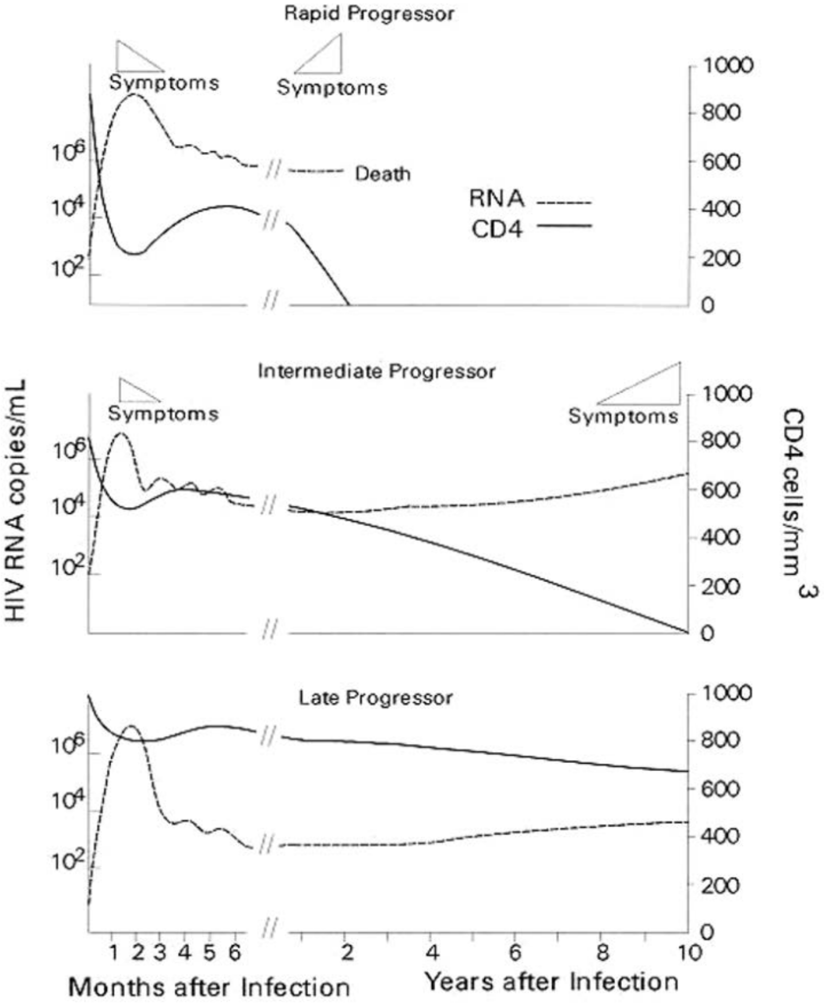

Figure 2-2: **Generalised comparison of different rates of HIV disease progression based on CD4 count and VL. Figure reproduced from Langford *et al.* (2007). (Used under fair dealing rights as described in the SA Copyright act)**

HIV-induced immune activation leads to the increased production and the secretion of pro-inflammatory cytokines causing inflammation and a persisting hypermetabolic state. During this hypermetabolic state, energy expenditure is up-regulated (Hommes et al., 1991). The up-regulation of energy would seem to be associated with increased mitochondrial function, as this is the production site of the bulk of the adenosine triphosphate (ATP). Williams (2012) however showed through the use of organic acids that HIV infection is associated with mitochondrial dysfunction. The NLRP3-inflammasome, which is a multiprotein complex that orchestrates innate immune responses to infection and cell stress, also link inflammatory changes to mitochondrial dysfunction (Aounallah et al., 2016). During HIV infection, the infected cells undergo metabolic reprogramming, resembling a Warburg-like effect, in an attempt to produce sufficient ATP through alternate means, i.e. namely through glycolysis (Aounallah et al., 2016). Several metabolites such as lipids, free fatty acids, dicarboxylic acids and Krebs cycle intermediates have thus been measured and found to be increased during HIV infection (Sitole et al., 2013). The interrelation between the immune and metabolic systems during HIV infection is evident.

2.2 HIV diagnosis and monitoring

Once infected, many individuals decide not to disclose their HIV status due to stigma. Stigma against HIV and AIDS is the standardised image of disgrace of infected individuals by the community at large. The UNAIDS' ambitious 90-90-90 target towards ending the AIDS pandemic calls for 90% of HIV-infected persons to know their status, 90% of infected persons to be on treatment and 90% of treated individuals to achieve viral suppression (Joint United Nations Programme on HIV/AIDS, 2014). Due to fear of stigma and discrimination, many choose not to test until they become sick enough to be compelled to do so. Delayed testing is the leading cause of late diagnosis and initiation of treatment (World Health Organisation, 2015). The choice not to disclose HIV status may hinder social and clinical support (Smith *et al.*, 2008). Although much effort is directed at destigmatising HIV, providing counselling and providing self-testing kits, it seems unlikely that the diagnosis target will be met by 2020.

During mid-2015, the World Health Organisation (WHO) published consolidated guidelines on HIV testing services (World Health Organisation, 2015). The main focus of this document was addressing the "5Cs", consent, confidentiality, counselling, correct results and connection. HIV testing services prescribed by the WHO comprises a full range of services that include pre- and post-test counselling, linkage to appropriate HIV prevention strategies, treatment and care, quality assurance and the delivery of correct results by laboratories (World Health Organisation, 2015).

The principles of the Enzyme-Linked Immunosorbent Assay (ELISA) or viral nucleic acid testing form the basis of most HIV tests. ELISA testing is based on the detection of an antigen (e.g. HIV p24 antigen and anti-gp120 antibodies) by a specific antibody coupled to an enzyme which, in the presence of a substrate, changes colour proportional to the amount of antigen. Viral nucleic acids can be detected early but involve much more intricate sample preparation procedures for analysis through PCR. PCR uses sequence-specific primers in conjunction with polymerase enzymes and nucleic acid substrates to target and amplify a specific region of DNA, in this case, a region of viral RNA reverse transcribed to DNA. During peak viremia, detection of HIV p24 antigen is possible. The detection of anti-HIV antibodies is dependent on seroconversion (time between anti-HIV antibody production and its detection in the blood).

The HIV Rapid test (also based on ELISA principles) is currently the most commonly used screening test due to its simplicity, low cost and rapid turnover time compared to other methods. Following a positive HIV diagnosis, a patient is linked to care and prevention services, provided with ART and monitored to confirm viral suppression. Two parameters most commonly used to monitor HIV infection and progression is the CD4 cell count which is measured in cells/µl blood, and VL represented as HIV RNA copies per millilitre plasma. Briefly, HIV infects CD4+ T cells which proliferate and undergo cell death either through apoptosis or cytotoxic attack. Since the number of CD4+ T cells and VL changes with the clinical presentation of individuals (i.e. changes throughout disease progression and at the initiation of anti-HIV treatment), these are the chosen parameters for monitoring HIV infection (Korenromp *et al.*, 2009). Even though CD4 count and VL are the preferred parameters, these are not without error. CD4 count in HIV-infected individuals, for example, can have a standard deviation of up to 26% per individual when monitored over time (Hughes *et al.*, 1994). Furthermore, inter-laboratory variation has also been reported as significant in independent tests in laboratories in Swaziland (Mlawanda *et al.*, 2012; Raboud *et al.*, 1995). Section 2.4 will direct attention to the longitudinal monitoring of CD4 and VL in light of HIV disease progression.

2.3 HIV treatment and vaccines

In prior years, being diagnosed with HIV was a death sentence. The unavailability of drugs and immune depletion, leading to AIDS, was the main reason for this sentence. Azidothymidine (AZT), a nucleoside reverse transcriptase inhibitor (NRTI) developed in the 1960s, was fortunately discovered to be a potent inhibitor of HIV replication. Over time the virus developed resistance against this treatment. This resistance was mainly due to several socio-economic factors that influence the coverage of and adherence to treatment. Viral resistance against drugs is common when treated with a single class of drugs, therefore the WHO suggests the use of combination antiretroviral therapy (cART) to combat the virus at several steps of its replication cycle (World Health Organization, 2013). This approach has led to the development of several different NRTIs and other drug classes to combat HIV. The most common drug classes additional to NRTIs are non-nucleoside reverse transcriptase inhibitors (NNRTIs), fusion inhibitors, integrase inhibitors and protease inhibitors (PI). In 1996, a combined treatment approach called highly active antiretroviral therapy (HAART) was introduced (Lange & Ananworanich, 2014). The U.S. Food and Drug Administration approved 40 antiretroviral medications belonging to the various classes up to March 2018 (U.S. Food and Drug Administration, 2018).

Before September 2016 (implementation of universal test and treat policy in South Africa), individuals were only issued with anti-HIV treatment when CD4 counts dropped below 350 cells/µl blood (World Health Organization, 2013). As of the 1st of September 2016 all HIV positive children, adolescents and adults in South Africa regardless of CD4 count were deemed eligible to be offered ART with a priority to those with CD4 counts ≤350 cells/ul blood (Department of Health, 2016). Modern treatment against HIV has increased the lifespan of individuals which displays phenotypically as delayed or slow disease progression.

Due to the highly polymorphic nature of HIV, vaccine development has to date been unsuccessful although it seems much more plausible than ten years ago. Klein *et al.* (2012) showed that broadly neutralising antibodies transferred to HIV-infected humanised mice were an effective controller of HIV-1 replication. Broadly neutralising antibodies have also shown excellent specificity and efficacy in binding to viral peptides and proteins *in vitro* (Sok & Burton, 2018). A new experimental HIV vaccine called “Mosaico” will enter stage III clinical testing in the last quarter of 2019. This vaccine targets more strains of HIV than any previously developed vaccine (Mega, 2019). Meanwhile in South Africa, as of February 2020, the HVTN 702 vaccine trial which had moved the furthest in human testing was stopped after it proved to be ineffective (UNAIDS, 2019).

While treatment has its benefits of prolonging patient survival, it also results in a myriad of metabolic complications and comorbidities in individuals. Understanding the underlying mechanism(s) of differential HIV disease progression may help identify protective factors to consider as part of treatment and vaccine design.

2.4 HIV disease progression

Since the discovery that HIV primarily infects CD4+ T cells, the primary biological marker for HIV infection has been CD4 cell count. Several other host and viral parameters, not of metabolic origin, were also tested for their use as markers of HIV infection and disease progression and included, for example, the measurement of: elevated serum β2 microglobulin and neopterin levels, levels of HIV p24, syncytium inducing HIV-1 phenotype, production of anti-HIV antibodies, etc. These markers lack sensitivity, specificity and predictive power, highlighting the need for less variable, earlier markers of disease progression (Gupta & Gupta, 2004). In subsequent years, the importance of longitudinal CD4 data was realised and appeared more in the literature (Lange *et al.*, 1992; Post *et al.*, 1996). Similarly, VL and longitudinal VL became synonymous markers for HIV prognosis and disease progression (Saag *et al.*, 1996). The availability of longitudinal data alongside biological samples primed researchers to investigate the various factors influencing HIV disease progression ranging from host to viral and environmental factors (Carrington & Walker, 2012; Chatterjee, 2010; Hahn *et al.*, 2018; Hazenberg *et al.*, 2003; Ipp *et al.*, 2014;

Langford *et al.*, 2007; Leserman, 2000; Leserman *et al.*, 1999; Leserman *et al.*, 2002; Scarpelini et al., 2016; Vujkovic-Cvijin *et al.*, 2013).

Principal host genetic factors impacting disease progression are those involved in the immune system. Of all the immunogenetic factors which influence HIV disease progression, the 32-basepair deletion in the *CCR5* gene is probably the most well-known biological change to confer protection against HIV infection (Huang *et al.*, 1996). This mutation translates to an altered *CCR5* protein which HIV needs as a co-receptor to fuse to the CD4+ T cell. Lack of the receptor ultimately leads to no or fewer infections of CD4+ T cells and a lower overall VL which ultimately translates to slower HIV disease progression. The mutation in the *CCR5* gene is not the only cause of slow disease progression and in context to our study has not been widely reported in individuals of African descent. However, not having the *CCR5* 32-bp deletion variant in African populations does not rule out a role for *CCR5* (other gene variants) in the context of our study.

Another significant parameter impacting HIV disease progression is the human leukocyte antigen (*HLA*) genotype (Carrington & Walker, 2012). The *HLA* gene cluster is located within the 6p21.3 region of chromosome 6 and contains more than 220 *HLA* genes (Anthony Nolan Research Institute, 2019). *HLA* genes are translated into cellular antigens that are involved in the identification of self and non-self cells (Carey *et al.*, 2019). There are two main classes of HLA molecules. Class I HLA molecules are expressed by all nucleated cells. Professional antigen-presenting cells present Class II HLA molecules. The *HLA* genes are highly polymorphic with more than 17 000 alleles in just the class I genes (Anthony Nolan Research Institute, 2019). Class I and class II molecules present intracellular peptides to CD8+ T cells and extracellular peptides to CD4+ T cells, respectively mediating cytotoxic and cell-mediated immunity depending the class of molecule presenting the antigen (Neefjes *et al.*, 2011).

Due to the genetic diversity of *HLA*, considerable heterogeneity exists between the efficiency of different alleles to bind specific processed peptides. In this regard, individuals that are homozygous for an allele have a reduced ability to present antigens and display a non-protective phenotype compared to the increased ability to bind antigens in heterozygous individuals who have a more protective phenotype (Fellay *et al.*, 2009). The *HLA-B* gene has more than 2000 alleles. Some alleles vary in the cytosolic region, and some vary in the extracellular region, while others vary in the peptide-binding cleft. It is for this reason that HLA, especially the *HLA-B* genotype, has a massive influence on whether HIV peptides will be bound and presented to the immune system hence impacting the host metabolism and patient outcome. Tumour cells and cells infected by viruses like HIV express foreign proteins which are presented to CD8+ T cells. If the CD8+ T cell receptor recognises the foreign peptide, a cytotoxic attack will commence on the presenting cell. Fast and non/slow progression is associated with several *HLA* alleles, notably *HLA-B*35px and *HLA-B*57, respectively (Miura *et al.*, 2008). HLA alone however does not fully explain differences in HIV disease progression (Olvera *et al.*, 2014). Brumme et al. (2009), for example, investigated the impact of selected immune and virological parameters on CD4 decline. High baseline CD4, low VL and protective *HLA* alleles correlated with a slower CD4 decline. The authors suggested combining VL and HLA markers to better define disease progression. In the study of Brumme *et al.* (2009), individuals with protective *HLA-B* alleles trend toward lower replication capacities of recombinant viruses encoding Gag-protease. The "unfit" virus thus translated to less viral replication and slower HIV disease progression (Wright *et al.*, 2010).

Doctor Shayne Loubser investigated the multiple roles of HLA in HIV immunity and treatment in a South African context during his doctoral degree (PhD) (Loubser, 2015). He identified vast differences in the *HLA-B* alleles and their frequencies among the various South African populations (Figure 2-3) which was later published (Loubser *et al.*, 2017). He showed that the previously identified KIR3DL1(-) KIR3DS1(+) *HLA-B* Bw4$^{80I/T}$ (+) protective haplotype had a higher prevalence among South African Indians and the lowest prevalence in the South African black population which partially explains the imbalance in HIV prevalence among South African populations. Other protective haplotypes had different distributions although the South African black population (which also lack the protective *CCR5* 32 basepair deletion) had an overall lower prevalence in protective haplotypes. This is important since samples from this study are from participants of black African descent.

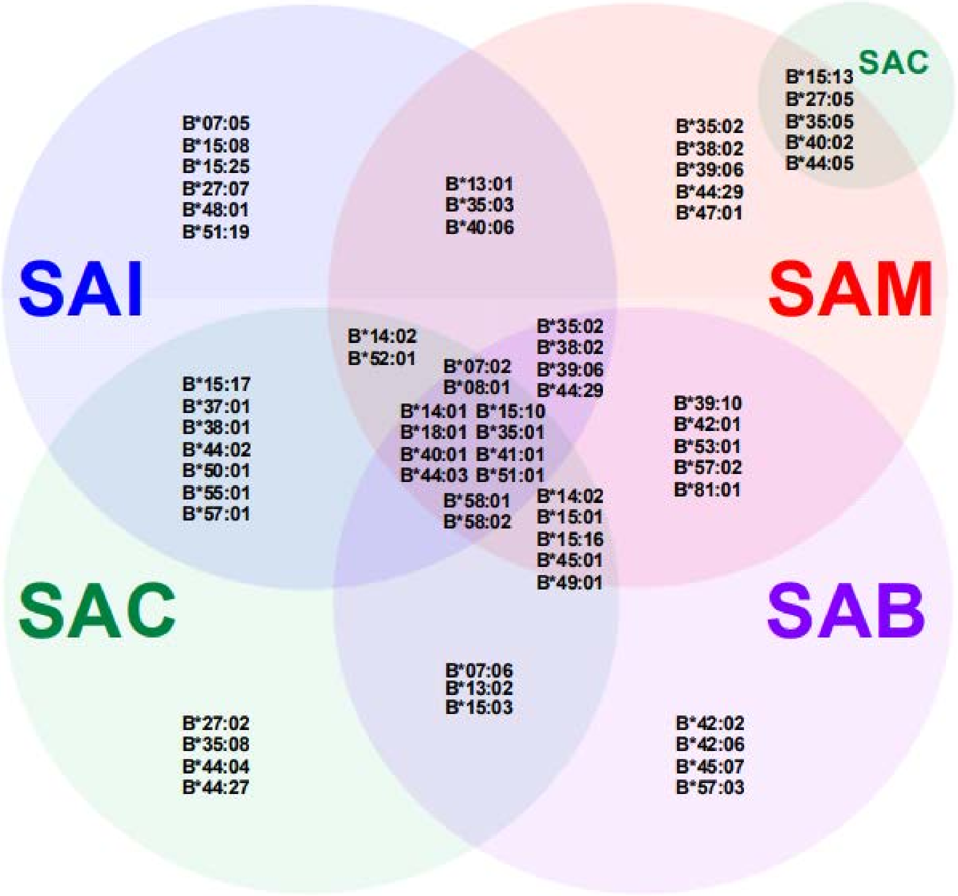

Figure 2-3: **Venn diagram showing the distribution of *HLA-B* alleles between South African Indian (SAI), South African Mixed Ancestry (SAM), South African Caucasian (SAC) and South African Black (SAB) populations. Figure reproduced from Loubser (2015) (Used under fair dealing rights as described in the SA Copyright act)**

As can be seen above, most literature report on immune and genetic factors leading to HIV control but these are not enough to explain the differential disease phenomenon. Therefore, continuous investigation of this field aims to understand the mechanisms of HIV control better. Miura *et al.* (2008) investigated the heterogeneity of virus found in elite controllers and found no common genetic defect in HIV coding genes that could explain HIV control. O'Connell *et al.* (2011) investigated the nature of CD4+ T cells from HIV-infected controllers and progressors. The cells of the controllers were more susceptible to HIV infection *in vitro*, which sounds controversial, but the faster rate of cell death in the cells of progressors explains the increased susceptibility to infection by the cells of the controllers. The cells of the controllers also produced less virus because they were less activated. The activation status of the cells thus correlated with virus production and informed on differential disease progression.

Other factors influencing progression is the subtype of the HI-virus. HIV subtypes have increased genetic diversity and display varying degrees of replication fitness which impacts on the patient's clinical phenotype. Venner *et al.* (2016) recruited and monitored Ugandan and Zimbabwean women newly infected with HIV-1 subtype C and D for clinical, social, behavioural, immunological and viral parameters over 3 to 9.5 years. A comparison between the Ugandan and Zimbabwean individuals showed the Zimbabwean women to progress slower. In both groups of women, those infected with HIV-1 subtype C progressed slower in the disease than those infected with HIV-1 subtype D. Subtype C compared to subtype D had lower replication capacity. This aspect may then hinder progression comparisons of samples within a cohort. Host genetics however differs between Ugandan and Zimbabwean individuals and may also account for the differences in HIV disease progression that were seen in this study.

Social, physiological and biochemical parameters also impact HIV disease progression. Faster disease progression has been associated with a stressful lifestyle (Leserman *et al.*, 2002), depression, lack of social support (Leserman, 2000), older age and even transmission between homosexuals (Langford *et al.*, 2007). These findings link HIV disease progression to more than just host immunity, viral and host genetic factors.

Although HIV disease progression is measurable through CD4 cell counts and VL, many researchers ignore the fact that these parameters have inherent flaws. These flaws include for example high variability which is as a result of the measured parameters' dependence on various host, viral and environmental factors. Many studies focus on elucidating the mechanism by which natural control of the virus is possible in these individuals, although the focus has mainly been on immune and genetic parameters. While several factors are associated with disease progression, the mechanisms of control of HIV infection is unclear, an aspect we try to address through a metabolomics-based approach.

2.5 Metabolomics

Metabolomics defines the unbiased identification and quantification of all intra- and extracellular metabolites in a biological sample using highly sensitive and selective analytical techniques (Dunn *et al.*, 2005; Dunn & Ellis, 2005). Metabolomics of blood-based biofluids gives a qualitative readout on intra-tissue "metabostasis" (homeostasis of metabolites) (Ivanisevic *et al.*, 2015). By comparing the metabolite profiles of samples with physiological differences (e.g. infection, time after an intervention, treatment etc.) biomarkers of the physiological variable can be detected (van Ravenzwaay *et al.*, 2007). Metabolomics makes use of techniques that chemically characterises biological samples. The chemistry of these biological samples varies significantly due to factors such as tissue type, genotype, gene expression, mutations, signalling molecules from other tissues or cells, availability of substrates for the various metabolic pathways, environmental factors and much more (Fiehn, 2002; Fiehn, 2016). Because metabolomics, especially untargeted metabolomics, aims to comprehensively characterise the chemical profile of biological samples, using metabolomics to investigate HIV disease progression will most certainly provide new insights to the chemistry of the process. Differential progression-guided clinical and immunogenetic changes can be measured in the metabolism, making metabolomics a suitable tool for such an investigation. While many metabolomics-based techniques are at our disposal, we elaborate on the application of GC-based systems for uncovering mechanisms relating to differential HIV disease progression.

2.5.1 Gas-chromatography mass-spectrometry (GC-MS)

Gas chromatography mass-spectrometry (GC-MS) techniques are often used in metabolomics studies. GC systems separate compounds by their distribution affinity between the liquid stationary-phase and gas mobile-phase as well as their volatility. The compound can only interact with the gas phase if it is in the gas phase itself. Therefore, a temperature ramp is used to sequentially evaporate the compounds so they can interact with the column. Two GC-MS analytical configurations are common, GC quadrupole MS and GC- triple quadrupole MS. GC-triple quadrupole systems typically ionise molecules in the source by chemical ionisation. A charge is transferred to the molecule by an ionised gas such as methane. The first mass filter is also most often a quadrupole followed by a collision cell and a second mass filter which can be either a quadrupole, ion trap or TOF, the separation which is performed by GC and is highly repeatable. The second and more common GC-MS configuration implements an ionisation source and a single mass filter. Electron impact ionisation (EI) ionises compounds in the source of the MS by bombarding eluting molecules with electrons. Although the ionisation energy (ionisation voltage) can be adjusted, the standard energy of -70 electron volts is used. Most commercial EI mass spectrometry (EI-MS) libraries are set up with ionisation energy of -70 electron volts. This standard configuration is a great advantage in compound identification for EI-MS implemented in GC-MS.

Another great advantage of GC is the repeatability of the separation. Since GC separation of compounds requires samples to be volatilised, salts, which only extract in trace quantities, do not volatilise at the normal operating temperatures of a GC and can therefore not interfere with ionisation by forming adducts. A disadvantage is that the maximum operating temperature of most GC columns is less than 350°C. Many polar compounds have a boiling point higher than this and have to be chemically modified by derivatisation to be compatible with GC-MS. Derivatisation is the chemical alteration of polar compounds by adding a non-polar functional group to the polar group. In reducing the overall polarity of the molecule derivitisation introduces many problems since it is not compound specific but rather functional group-specific and will derivatise most or all compounds which have a specific polar functional group. Silanation is a preferred method for derivatising samples. During derivatisation, a trimethyl silane group covalently bonds to hydroxy, amine and some carbonyl groups. Unfortunately, various factors influence the derivatisation of specific groups such as stereochemistry and other derivatised groups, leading to a heterogeneous mixture of the derivatised product from a single molecule (Knapp, 1979). Carbohydrates, which have many hydroxy groups, can have up to 16 different derivatised products, thereby increasing the complexity of the sample. Additionally, carbonyl groups do not generally derivatise first, and is stereochemically affected by the derivatisation of the more preferred groups such as alcohols, phenols, carboxylic acids, amines and amides (Knapp, 1979). This leads to a heterogeneous mixture of derivatised and underivatised carbonyl groups from a single molecule containing more than one carbonyl group. Since derivatisation is a dynamic process in which the ratio between derivatised products from a single source changes over time and time between derivatisation and analysis is not constant for all samples, time after derivatisation introduces a major source of variation. Effectively, samples analysed immediately after derivatisation will have different ratios of derivatised products to a sample analysed several hours later (e.g. at the end of a batch). Minimising batch sizes and randomising the samples injection order effectively reduces the time variation.

One of the biggest advantages of MS-based systems is sensitivity. Sensitivity will be crucial in obtaining a comprehensive metabolic profile of HIV-infected samples with which to characterise HIV disease progression. Because compounds are separated and their spectra analysed sequentially, compounds in low concentrations reach the detector separated and purer, which means that they can be better identified and quantified.

2.5.2 2D-GC techniques

The human metabolome database (HMDB) contains entries for more than 110 000 metabolites (Wishart *et al.*, 2017). This vast number of detectable metabolites is attributed to the popularity of targeted metabolomics. In targeted metabolomics the sample matrix is simplified by extraction of relevant metabolites or the analytical platform is set up to only quantify a preset list of compounds or a combination of these (Roberts *et al.*, 2012). During untargeted metabolomics approaches, however, the aim is to identify and quantify as many metabolites present in a sample. Unfortunately, most sample matrices are not suitable for direct analysis. The extraction of metabolites thus becomes crucial (Yu *et al.*, 2017). Conventional GC-MS systems have a peak capacity of about 400 peaks while modern "super-resolution" GCxGC platforms can distinguish up to 10,000 peaks (Liu & Phillips, 1991; Marriott & Nolvachai, 2018). This massive increase in peak capacity allows researchers to minimise sample preparation yet still resolve the peaks from as many compounds present in detectable concentrations.

Two-dimensional gas-chromatography coupled to time of flight mass spectrometry (GCxGC-TOFMS) relies on the same principles for separation of compounds as conventional GC-MS, namely volatility and column interactions (polarity). What sets two-dimensional gas-chromatography (2D-GC) apart from conventional systems is the addition of a second column with a different stationary phase which separates most co-eluting compounds through the different chemical properties of the stationary phase of the second column. Typically, conventional GC systems, as well as 2D-GC systems, use a low-polarity Rxi-5sil-MS column with a cross bond 1,4-bis(dimethylsiloxy)phenylene-dimethyl-polysiloxane stationary phase as the primary column in untargeted metabolomics analysis. The secondary column typically used is a mid-polar Rxi-17 column with a cross-bond diphenyl-dimethyl-polysiloxane stationary phase. The low polarity column will have a higher affinity for apolar compounds and thus retain them longer, thus eluting them later in the chromatogram. The low polarity primary column is ideal for GC analysis since small apolar compounds generally have lower volatilities and are thus more suitable for GC analysis than larger polar molecules with high volatilities. Derivatisation reduces the polarity of these large polar molecules by masking the electronegative groups. Low polarity compounds have more interactions with the apolar primary column and are thus separated better in the first column. Since heterogeneity still exists in the polarity of compounds, more polar compounds will have fewer interactions with the column and will thus not separate as well as apolar compounds. The increased polarity of the second column allows for more interactions between these polar metabolites and stationary phase, thus also allowing for better separation of polar compounds.

Conceptually, 2D-GC has been around for many years, but interfacing between the two separating dimensions while maintaining the resolution of the first dimension was a problem. Apolar compounds with different volatilities will separate in the first column, but due to the higher retention in the second column, these compounds might co-elute from the second column. A catch and release mechanism had to be designed to keep compounds that separated in the first column apart in the second column. “Heartcutting” was the first solution whereby a fraction of eluent from the first column was captured and reanalysed on a second column (MacNamara *et al.*, 2004). “Heartcutting” was inefficient as the second dimension data was only available for a small retention time window of the first dimension. The concept of capturing eluents from the first dimension and “injecting” them into the second dimension inspired the development of thermal modulation systems. Thermal modulation works by trapping eluents in a narrow band inside the column by condensing them in a supercooled segment of the column but still allowing the carrier gas to pass through (Liu & Phillips, 1991). Thermal modulation allowed multiple shorter “heartcuts” over the entire run. Thermal modulation consists of two sets of orthogonal hot and cold air jets that each blow over a small (1cm) section of the column. The double set of jets allows for continuous trapping of first dimension eluent and rapid release of trapped eluent from the modulator into the second column. Cold air for the trapping jets is obtained either through refrigeration at -80°C or cooling with liquid nitrogen. Thermal modulation systems have paved the way for comprehensive 2D-GC analysis with super resolution.

Although conventional GC systems coupled to sensitive time of flight mass spectrometers (TOFMS) exist, they don’t provide as much of an analytical advantage as 2D-GC systems coupled to TOFMS. What sets TOFMS apart from normal quadrupole systems is their sensitivity. Quadrupoles act as mass filters scanning through the ions removing everything except for the ion in the current scan which if present will produce a signal at the detector proportional to the number of ions. Each full scan represents a spectrum showing the intensities of all scanned ions for that scan. In untargeted metabolomics investigations, ions from 50 Daltons up to 800 Daltons are of interest, and therefore the scan range is set accordingly. Because quadrupoles scan through these masses (assuming the charge is +1), only 1/750th of the time of a scan is allocated to each mass. The scanning nature of quadrupoles means that 749/750 of the time each ion is filtered and does not reach the detector.

In contrast, in TOFMS systems, the ionised fragments are briefly trapped then “shot” with an electromagnetic pulse toward the detector. The energy transferred to each fragment is proportional to its charge. Assuming all fragments have a charge of +1, they are all accelerated through a vacuum to the detector with the same force. According to Newton’s second law, the acceleration is equal to the product of the mass of, and the force on the accelerating object. In the case of TOFMS, the objects are the fragments. Fragments with higher mass will thus accelerate slower than lighter fragments with the same charge. When the electromagnetic pulse ends, these fragments move at different constant speeds inversely proportional to their mass- to-charge ratio. At these different speeds, the time taken to reach the detector is proportional to the speed. The higher the speed, the faster the time to the detector. The lighter, faster-moving fragments reach the detector first, and the heavier, slower-moving fragments reach it last. The time from the pulse to detection is thus proportional to the mass-to-charge ratio of the fragment. By design, all fragments reach the detector giving TOFMS much higher sensitivity than quadrupole MS systems.

Additionally, the scan rate of quadrupoles is much slower at around 20 scans per second compared to the 1000 pulses per second in a TOFMS. For this reason, the number of data points over a peak is much higher in TOFMS systems than in quadrupole systems leading to better peak finding and more accurate area calculation (Dettmer *et al.*, 2007).

2D-GC systems, coupled to a TOFMS combines the super-resolution of the GCxGC separation with the sensitivity and scan rate of the TOFMS to provide a truly remarkable analytical platform.

2.6 Non-metabolomics investigations of HIV infection and disease progression

Various biological associations to HIV infection have been identified using relatively basic assays or assays not targeted to metabolites (e.g. ELISA). Lipodystrophy, hyperlipidemia, insulin resistance, diabetes, inflammation markers, coagulation markers, immune dysfunction, immune activation and various cytokines (Graber, 2001; Kuller *et al.*, 2008; Liovat *et al.*, 2012; Nixon & Landay, 2010; Williams *et al.*, 2013) have all been associated with HIV infection.

HIV primarily infects cells of the immune system and destroys them either by infection-induced apoptosis or killing by cytotoxic immune cells. The release of cytokines during this process serves as a communication medium between cells of the immune system. On first thought, this may not justify a metabolomics analysis of HIV, but HIV-induced metabolic reprogramming as well as the systemic effects of HIV infection through inflammation and immune suppression affects various organs and tissues and thereby also the metabolism of these organs and tissues leading to a phenotypical hypermetabolic state (Aounallah et al., 2016; Hommes et al., 1991; Lake & Currier, 2013; Palmer et al., 2014; Vassimon et al., 2012). The up-regulation of energy production and use thereof was suggested to occur via elevated mitochondrial function, as this is where the bulk of adenosine triphosphate (ATP) production occurs.

Several studies show that changes in either the immune system or the metabolism directly affects the other (Aounallah et al., 2016; Dagenais-Lussier et al., 2015; Fitzpatrick & Young, 2013; Kapoor et al., 2012; O'Neill et al., 2016). Interferon-gamma (IFN-γ), interleukin(IL)-1, IL-6 and tumor necrosis factor-alpha (TNF-α) are pro-inflammatory cytokines, and their presence has been shown to link the inflammatory response to metabolic changes (Kapoor et al., 2012). Using a cytokinomics approach, Williams et al. (2013) showed elevated pro-inflammatory IL-6 and IL-10, to discriminate between patients at an advanced stage of HIV infection versus those who were not. These cytokines are associated with specific metabolic changes. Lipid metabolism, glucose management and mitochondrial impairments have also been associated with inflammation and metabolic change (Aounallah et al., 2016; Williams, 2012). Immune responses characterised by changes in IL-1, IL-6 and TNF-α stimulate leptin production, decrease lipase activity and decrease adipose glucose uptake, causing hypertriglyceridemia (Slama et al., 2009). These cytokines, as well as hypertriglyceridemia, are reported during HIV infection. The oxidation of free fatty acids during HIV infection is increased to keep up with increased resting energy metabolism and to decrease hypertriglyceridemia (Fitzpatrick & Young, 2013; Hommes et al., 1991). Hyperlipidaemia, lipodystrophy, insulin resistance and diabetes represent some of the clinical symptoms of this hypermetabolic state induced by HIV infection (Graber, 2001). Consequently, several metabolites influencing HIV disease progression have been identified. Cholesterol and tryptophan detected in conventional assays are examples of these.

Tryptophan can be catabolized to kynurenine by indoleamine 2,3-dioxygenase (IDO) or to serotonin by tryptophan 2,3 dioxygenase (TDO). The expression of IDO/TDO is generally increased during HIV-induced immune activation. The kynurenine produced as a result of tryptophan catabolism has immunosuppressive effects which affect disease progression in individuals. Kynurenine generally reduces IL-2 signalling (Vesterbacka *et al.*, 2017). Jenabian *et al.* (2013) investigated the catabolism of tryptophan and its impact on the T helper 17 and T regulatory cell balance using online solid-phase extraction LC tandem MS. HIV is known to decrease T helper 17 and increase T regulatory cell populations resulting in altered mucosal immunity, translocation of microbes from the gut and further immune activation (Paiardini *et al.*, 2008). A comparative analysis between uninfected healthy participants and populations treated and untreated for HIV as well as natural controllers of HIV infection showed the controllers to present with reduced expression of TDO, reduced levels of kynurenine-based metabolites, increased frequency of Th17 cells and low levels of inflammatory IL-6 (Jenabian *et al.*, 2013). IDO-induced immunometabolism is therefore suggested as an inflammation-related marker for HIV disease progression. A study by Cassol et al. (2013) showed that ART-naïve participants had lower levels of tryptophan that those on treatment. They also showed that elite controllers (EC) had significantly lower levels of kynurenine compared to progressors. The reduced kynurenine suggests that ECs have unique tryptophan catabolism, confirmed by decreased IDO mRNA expression.

A disturbed gut microbiome is generally associated with HIV disease progression. The alteration in gut mucosal immunity in HIV-infected patients was studied by Vesterbacka et al. (2017). The microbiome biodiversity in ECs was much richer than that of uninfected participants. This rich biodiversity is likely due to an alteration of the mucosal immunity because even though the virus is suppressed through medication, immune activation still persists. In the ECs the authors also reported an inverse correlation between CD4 count and carbohydrate metabolism as well as a positive correlation between CD4/CD8 T cell ratio and degradation of valine, leucine and isoleucine.

In 1991, one of the first research papers linking cholesterol to HIV infection was published. The results showed a positive correlation between CD4 count and low-density lipoprotein (LDL) cholesterol and that this could be used as a discriminating parameter for discerning the progression of a patient (Rübsamen-Waigmann et al., 1991). This was later confirmed by (Fletcher et al., 1993). Trans-infection occurs when the HI-virus is transmitted between cells during the interaction of HLA molecules with T-cell receptors. Rappocciolo *et al.* (2014) studied trans-infection between antigen-presenting cells (dendritic and B cells) and T cells. The authors found that reduced levels of cholesterol in the antigen-presenting cells associated with reduced trans-infection. This association is most likely attributed to the role of cholesterol in cell signalling, adhesion, permeability and membrane integrity (little or no synapses for HIV transfer).

Given the insights gained from studying HIV-induced metabolic parameters, it made sense that researchers would soon want to investigate more holistic changes hence developing metabolomics-approaches for this purpose.

2.7 Metabolomics investigations of HIV infection and disease progression

Early non-metabolomics investigations of HIV revealed its effect on the nervous system (Merrill & Chen, 1991), glucose homeostasis (Liegl, 1995) and the gut (Lu & Kotler, 1995; Quesnel et al., 1994; Reka & Kotler, 1990; Zeitz et al., 1992). Later, more metabolite focussed studies identified lipid abnormalities, insulin resistance and diabetes in HIV-infected treated individuals (Graber, 2001). Hewer et al. (2006) first applied metabolomics to characterise HIV/AIDS and its associated treatment profiles. Metabolomics was later applied in 2011 for the characterisation of the disease by using oral metabolites of infected versus non-infected participants (Ghannoum et al., 2011; Maher et al., 2011). Since then, metabolomics has been used to characterize HIV in various biofluids including bronchoalveolar lavage (Cribbs et al., 2014; Cribbs et al., 2016), cerebrospinal fluid (Cassol et al., 2014; Dickens et al., 2015; Wikoff et al., 2008) and plasma (Cassol et al., 2013; Hodgson et al., 2017; Scarpelini et al., 2016; Sitole et al., 2013).

In the case of HIV, metabolomics has already been used to differentiate HIV positive and negative samples from one another (Armah *et al.*, 2012; Chetwynd *et al.*, 2017; Dagenais-Lussier *et al.*, 2015; Epstein *et al.*, 2013; Hollenbaugh *et al.*, 2011; Kuller *et al.*, 2008; Langkilde *et al.*, 2015; McKnight *et al.*, 2014; Nonodi & Meyer, 2014; Palmer *et al.*, 2014; Philippeos *et al.*, 2009; Rodríguez-Gallego *et al.*, 2018a; Sitole *et al.*, 2014; Sitole *et al.*, 2015; Swanson *et al.*, 2009; Tarancon-Diez *et al.*, 2019; Vázquez-Castellanos *et al.*, 2018; Williams *et al.*, 2013; Zhang *et al.*, 2018). The approach has also been used to differentiate HIV positive individuals from those treated with ART (Cassol *et al.*, 2013; Graber, 2001; Justice *et al.*, 2010; Koethe *et al.*, 2016; Moutloatse *et al.*, 2016; Mulligan *et al.*, 2000; Munshi *et al.*, 2013). The neurological effects of HIV and AIDS have also been studied using metabolomics (Cassol *et al.*, 2014; Dickens *et al.*, 2015; Pendyala & Fox, 2010; Wikoff *et al.*, 2008) as well as associations with alterations in the gut (Vázquez-Castellanos *et al.*, 2018).

HIV-induced metabolic reprogramming mainly alters glucose metabolism. HIV infection is associated with an increase in the expression of glucose receptors and the levels of glycolytic intermediates suggestive of elevated glucose metabolism (Ahmed *et al.*, 2018). TNF-α and IL-6 levels generally increase with elevated glucose metabolism. In immunological non-responders (progressors) this upregulation persists. In a non-metabolomics study, Hegedus *et al.* (2014) showed that galactose supported the growth of cells and reduced virus replication. This result shows that alternative carbohydrate energy sources mediate antiviral responses. Metabolomics has also been applied to study the effects of treatment in HIV-infected individuals. Results similar to those of conventional techniques were initially seen, such as insulin resistance and hyperlipidaemia (Graber, 2001; Mulligan et al., 2000). Munshi et al. (2013) used a nuclear magnetic resonance (NMR) approach to investigate the metabolic differences between the plasma, urine and saliva of HIV negative controls, HIV positive individuals as well as HIV positive individuals on treatment. The authors found that changes in plasma metabolites identified with NMR were the most significant. In plasma, they found an increase in aspartic acid, glucose, sarcosine, methylmalonic acid and choline in the HIV positive group but a significant decrease of these molecules in ART receiving participants. Fumaric acid increased in HIV positive participants as well as the ART receiving participants. Acetoacetate and threonine decreased in HIV positive participants with slightly higher levels in the treated participants, but still less than the control group. Similar to previous reports, amino acid biosynthesis and carbohydrate metabolism were differentially regulated (Cassol *et al.*, 2013; Munshi *et al.*, 2013). Cassol *et al.* (2013) additionally linked lipid abnormalities found in HIV positive participants to microbial translocation and hepatic function when treated with protease inhibitors. An NMR metabolomics study done on the effects of ART on neonates revealed increased 3-hydroxy butyric acid indicating a combined failure in lipolysis and ketogenesis as well as hypoxanthine which is a proposed indicator of hypoxia of newborns (Moutloatse *et al.*, 2016). Metabolomics, therefore, has shown in various biofluids and cohorts that significant metabolic differences exist between uninfected and HIV-infected individuals that are untreated and treated.

As part of the HIV-based metabolomics investigations, various platforms have been utilised for the analysis of samples namely; NMR, Raman spectroscopy, LC-MS, attenuated total reflectance Fourier-transform (ATR-FTIR) and nanoflow nanospray mass spectrometry to name but a few (Chetwynd *et al.*, 2017; Maher *et al.*, 2011; McKnight *et al.*, 2014; Munshi *et al.*, 2013; Nonodi & Meyer, 2014; Philippeos *et al.*, 2009; Sitole *et al.*, 2014; Sitole *et al.*, 2015; Swanson *et al.*, 2009).

Although GC methods have been utilised to characterise HIV infection (Sitole *et al.*, 2013; Tarancon-Diez *et al.*, 2019), untargeted 2D-GC analysis of HIV-infected samples has not been reported.

Metabolomics studies which focus on characterising HIV disease progression have received little attention. Although many genetic, immune and proteomic links with HIV disease progression exists, metabolomic links on the topic are lagging. In fact, the first metabolomics article on HIV infection was published only recently (Hewer *et al.*, 2006). In context to applying a metabolomics-based approach to studying HIV disease progression, Scarpelini *et al.* (2016) were one of the first to investigate the plasma metabolite profile of HIV-infected participants at different stages of disease, different pace of progression, different viremic levels and immunologic response to treatment. Since metabolomic investigations pertaining to disease progression is more commonly applied now, post-adoption of the UTT policy, researchers have come up with additional terms for defining HIV disease progression in the context of treated individuals. In these studies, response to treatment is used to define groups displaying an improved (non-progressive) or worsened outcome (progressive). In the study of Scarpelini and colleagues, the authors followed a targeted LC-MS approach for 186 metabolites including acylcarnitines, amino acids, amines, hexoses, phosphatidylcholines and sphingomyelins. They identified five metabolites which distinguished rapid progressors and immunological non-responders at baseline. The metabolites identified to differentiate rapid progressors, and immunological non-responders from non-progressors and immunological responders were linked to acylcarnitine metabolism as well as the catabolism of lysine, organic acids and tryptophan, which were all increased in EC. The authors identified severe deregulation in sphingomyelin and acylcarnitine metabolism and suggested that ECs of HIV might harbour late-onset multiple acyl-coenzyme A dehydrogenase deficiency (MADD) which represents a defect in the body’s ability to process fats to produce energy. They hypothesise that an asymptomatic inborn error of metabolism, MADD, relates to control of HIV. Subsequent enzyme activity tests on electron-transferring flavoprotein dehydrogenase, an enzyme associated with MADD, showed that the enzyme activity is significantly lower in ECs compared to rapid progressors (Scarpelini et al., 2016).

In a follow-up article, the authors performed HLA typing and analysed the gene expression profile of individuals presenting the non-progressor, rapid-progressor, immunological responder and immunological non-responder clinical phenotypes, respectively (Zanoni *et al.*, 2017). Elite controllers presented with protective *HLA* alleles, had unique transcripts associated with presentation between DCs and antigen-specific CD8+ T cells, expressed cytokines with anti-inflammatory properties and mRNA with stem cell mobilisation. Although their studies were successful in identifying metabolite markers linked to disease progression, larger sample groups and an untargeted metabolomics approach might yield more discerning metabolites which may lead to discoveries beyond the 186 metabolites targeted by these authors.

In the study of Rodríguez-Gallego *et al.* (2018c), treated individuals with an improved immunological outcome (slower progression) presented with increased high density lipoprotein, total cholesterol, branched-chain amino acids and tyrosine. High density lipoprotein has a role in innate and adaptive immunity with its associated proteins impairing HIV fusion and cell penetration. Those individuals with a worsened outcome presented with an increase in glycolytic intermediates, LDL and very low-density lipoprotein.

Recognising the wealth of information that can be obtained through the combined measurement of immune and metabolic parameters, Tarancon-Diez *et al.* (2019) investigated mechanisms that differentiate those individuals who spontaneously lose virological control (progressive group) over those who maintain it (non-progressive group). The authors employed untargeted and targeted GCqTOFMS and LCqTOFMS approaches. Before the loss of virological control, the infected individuals presented with aerobic glycolysis, mitochondrial dysfunction, immune activation and increased oxidative stress. The discriminatory metabolite was the amino acid valine, increased in the progressive group. These changes associated with a decrease in polyfunctional HIV-specific CD8 responses.

Taken together, the imperfections of current prognostic indicators, as well as the lack of metabolic indicators for defining HIV disease progression, is evident. There remains a gap in understanding how various clinical factors and more especially immunogenes impact on metabolism and ultimately the host phenotype (in this case, HIV disease progression). This gap can potentially be addressed through metabolomics which measures the metabolic responses of living systems to biological stimuli (Kamleh *et al.*, 2009). While genetic studies are prominent, these can only identify the genes responsible for the phenotype, not how the gene exerts its function to cause the phenotypic variation in progression. Proteomic functional studies can study single interactions between proteins but not the full metabolomic effect. Because metabolites are the products of the transcriptome and the proteome, metabolomics provides an improved approach with which to study the metabolic reprogramming caused by HIV-induced immune activation. An untargeted metabolomics approach may reveal the downstream effects of these clinical parameters and immunogenes that predict HIV disease progression. Two-dimensional gas-chromatography coupled to a time-of-flight mass spectrometer (GCxGC-TOFMS) is an especially suited analytical platform for analysing such chemically complex samples with high sample dimensionality (Schoeman & Loots, 2011). Although not the primary focus of this study, we investigate in addition to the metabolic changes, the cytokine profile of the samples to explore associations between the immune and metabolic systems in the context of our cohort. Preliminary findings are shown in Appendix I).

SAMPLING AND DESIGN

3.1 Introduction

The time constraints linked to an M.Sc. study removes the possibility to recruit and monitor individuals over time. Furthermore, South Africa has implemented a test and treat policy, which makes a study on untreated participants unethical. Due to these limitations, we could only use previously collected samples. The primary focus of this study was to characterise HIV disease progression. Progression inherently adds a time factor to the study. In the HIV context, clinical progression can only be established after many years of measuring CD4 counts and VL, thus encompassing the aspect of monitoring. CD4 count and VL are established prognostic indicators of HIV infection. Although they are routinely used in the clinic, these parameters vary quite significantly (Korenromp *et al.*, 2009). Many data points are thus needed to be confident that CD4 count and VL are either increasing or decreasing. The metabolomics approach adds its own set of criteria concerning the analysis of the sample. Many factors such as age, gender, ethnicity, lifestyle, medication etc. are known to influence the human metabolome (Johnson & Gonzalez, 2012). These factors imply that a metabolomics approach can only be efficient in a homogenous group, with the factor being analysed (progression) as the main difference between compared groups.

3.2 Overview of the experimental design

The experimental design comprised the selection of the participants. Participants were selected by applying inclusion and exclusion criteria filters. This was followed by sample selection, extraction, analysis as well as statistical analysis of the data. Sample selection was complex for this specific project since selection criteria were strict and the option of recruiting new participants eliminated. As previously mentioned, the test and treat policy adopted in South Africa makes it unethical to include participants in a study without providing them with treatment. Being on treatment formed part of the exclusion criterion and therefore, new samples for a metabolomics investigation of HIV disease progression was not acquired. Therefore, previously collected blood plasma samples had to be used. Fortunately, Prof Thumbi Ndung'u from the University of KwaZulu-Natal (UKZN) agreed to share blood plasma previously collected from an untreated HIV positive population that had the relevant longitudinal CD4 count and VL previously analysed, in order to determine progression, as well as the associated HLA genotype data. In order to classify participants as either progressing or non/slow-progressing, CD4 cell counts had to be determined in these HIV positive individuals for a period of at least four years. The average change in CD4 cell counts over this period were used to classify participants as either progressors or non/slow-progressors. As part of the initial study, the untreated HIV participants' blood plasma were also

HLA-typed. This allowed for classification of the HIV positive individuals as presenting either protective or non-protective phenotypes. Classifications were based on HLA alleles that were previously linked in the literature with phenotypes of HIV disease progression. Plasma samples were selected considering only the CD4 counts at a specific time point. For sub-study 1, CD4 cell count is the independent variable and plasma samples with high (500-800 cells/µl) and low (<250 cells/ µl) CD4 counts were chosen and compared. Time points were selected for plasma samples with median CD4 cell counts (350-499 cells/ µl) to allow for the comparison of the sample groups of sub-study 2 and 3.

Once the samples were selected, they were requested from the UKZN. At the UKZN, the collected plasma samples were stored in access-controlled, and temperature monitored freezers at -80°C. The samples were transported to the North-West University (NWU) on dry ice, thawed and aliquoted on arrival. Sample aliquots of 50 µl, as well as the pooled QC samples, were stored at -80°C until the day of the extraction.

Standardisation of the metabolomics methodology, as is explained in 4.4.1., was accomplished using the QC samples. On the day of metabolite extraction, a single QC sample and a batch of randomised samples were thawed, extracted and derivatised according to the in- house standard operating procedure (SOP) described in 4.3.3. After extraction, the samples were analysed on the Leco Pegasus 4D (GCxGC-TOFMS) system using the standardised method, as explained in 4.4. After all the batches were analysed, the data was processed using the vendor-specific software, ChromaTof (version 4.72) and its associated statistical compare package, to find peaks, identify the compounds and align the peak data. The peak areas were exported for subsequent statistical analysis.

Statistical analysis was performed on the data as per the respective sub-studies. The samples pertaining to each sub-study were stratified and analysed independently. For sub-study 1, all samples in the mid-CD4 group were excluded, and only the high- and low-CD4 groups compared. For sub-study 2 and 3, the high- and low-CD4 groups were excluded, and only samples from the mid-CD4 group used to compare the groups. The samples from the mid-CD4 group were stratified according to their change in CD4 over time (sub-study 2) and *HLA-B* alleles (sub-study 3), respectively. The metabolite profiles of the mid-CD4 group were subsequently statistically compared based on these classifications.

The purpose of the statistical procedures used for all three sub-studies (e.g. p-value<0.05, effect size > 0.5 etc.), was to select those metabolites varying the most or best explaining the variation seen when comparing the groups in each of the respective sub-studies. Figure 3-1 shows a graphical summary of the experimental design.

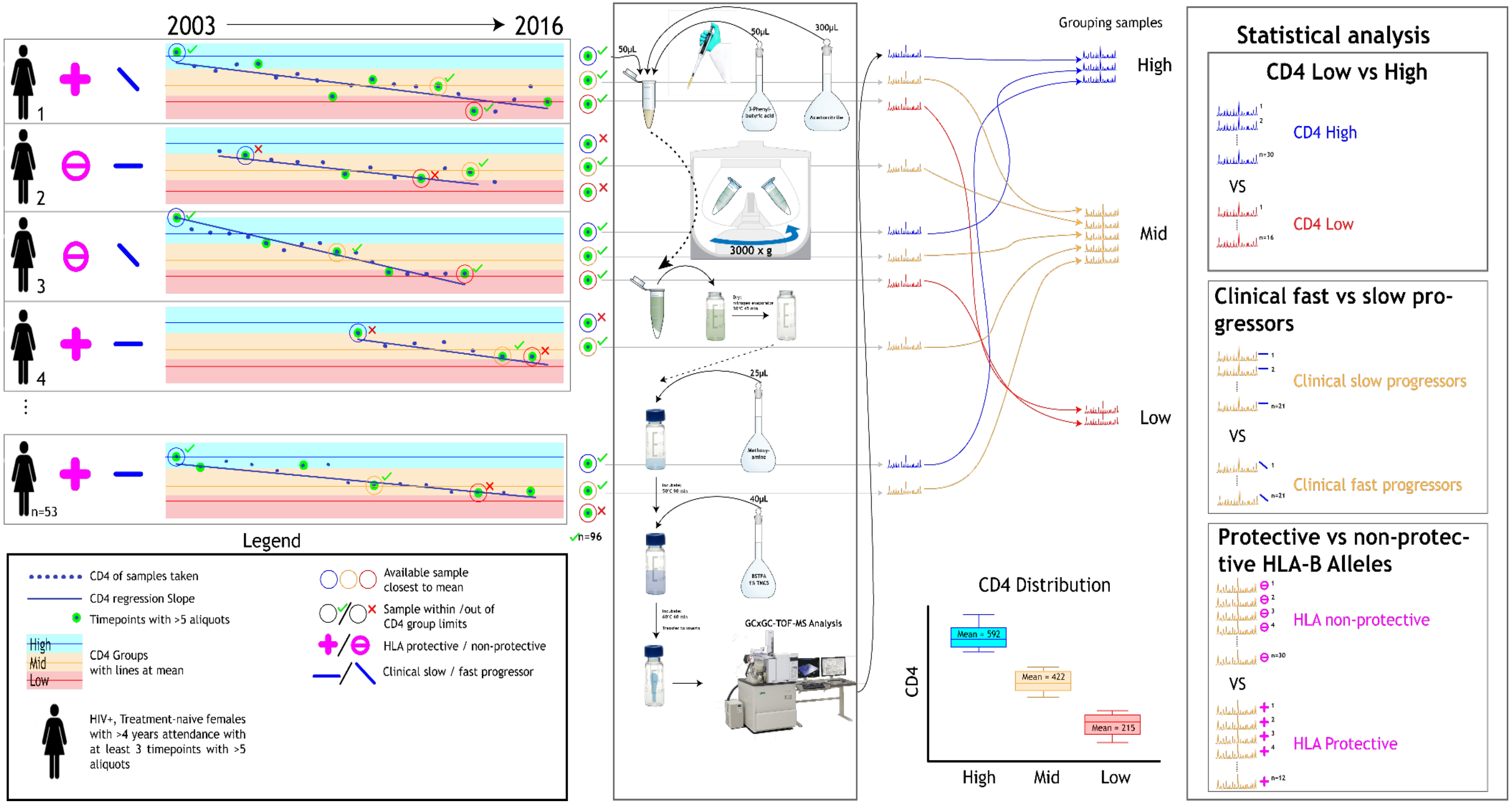
2003
2016
1
2
3
4
⋮
n=53
n=96
Legend
CD4 of samples taken
CD4 regression Slope
Timepoints with >5 aliquots
High
Mid
Low
CD4 Groups with lines at mean
HIV+, Treatment-naive females with >4 years attendance with at least 3 timepoints with >5 aliquots
Available sample closest to mean
Sample within /out of CD4 group limits
HLA protective / non-protective
Clinical slow / fast progressor
50µL
50µL
300µL
3-Phenyl-butyric acid
Acetonitrile
3000 x g
Dry: nitrogen evaporator 30°C 45 min
25µL
Methoxy-amine
Incubate: 50°C 90 min
40µL
BSTFA 1% TMCS
Incubate: 60°C 60 min Transfer to inserts
GCxGC-TOF-MS Analysis
Grouping samples
High
Mid
Low
CD4 Distribution
CD4
Mean = 592
Mean = 422
Mean = 215
High
Mid
Low
Statistical analysis
CD4 Low vs High
CD4 High
n=30
VS
CD4 Low
n=16
Clinical fast vs slow pro-gressors
Clinical slow progressors
n=21
VS
Clinical fast progressors
n=21
Protective vs non-protec-tive HLA-B Alleles
HLA non-protective
n=30
VS
HLA Protective
n=12

Figure 3-1: Summarised experimental design showing the selection of participants and their respective samples as well as the extraction of these samples for analysis on a GCxGC-TOFMS system. Stratification of the samples based on CD4 count, progression status and *HLA-B* alleles allowed for statistical comparisons in line with the respective sub-studies.

3.3 Collaboration with UKZN and background to the samples

As this study was time-limited, a collaboration was established with the HIV Pathogenesis Programme (HPP) of the University of KwaZulu-Natal. HPP had samples of a cohort of HIV patients that had been followed longitudinally and that was well-characterized clinically (Brumme *et al.*, 2009). The HPP agreed to collaborate by providing us with the samples and relevant longitudinal data. The samples were initially collected as part of a study titled: "Characterisation of the evolution of adaptive and innate immune responses in HIV clade C virus infection". The aim of the initial study was to gain a better understanding of HIV disease progression from an immunological standpoint in the Sinikithemba cohort. The department of Health, KwaZulu-Natal, granted permission to conduct the respective studies. Site approvals were signed off by managers at the respective clinic/hospital sites. These studies were approved by UKZN's biomedical research ethics committee (BREC) (E028/99 and E036/06). Since the immune and metabolic systems are interlinked, our work set out to understand HIV disease progression from a metabolic standpoint and therefore complements the aims of the initial study. This study was also approved by the Health Research Ethics Committee (HREC) of the Faculty of Health Sciences at the NWU (reference number NWU-00125-17-A1).

Since this was a pilot study, it was agreed that 100 samples be analysed for the purposes of the current study. The criteria for participant and sample selection is discussed below (Sections 3.4 and 3.5, respectively).

3.3.1 Plasma collection protocol

As per the protocols supplied to us, whole blood was collected from untreated HIV positive individuals regularly between 2003 and 2016. Utilising standard procedures, plasma was prepared from blood collected in EDTA tubes. These samples were aliquoted and stored at -80ºC.

3.3.1.1 CD4 and VL determination

CD4 cell count was enumerated from whole blood using the TruCOUNT method on the FACSCalibur instrument (BD Biosciences, San Jose, USA). The VL was determined with the Roche amplicor assay, version 1.5 or Cobas TaqMan HIV-1 test.

3.3.1.2 *HLA* determination

HLA typing was initially performed on an oligo-allelic level with the Dynal RELITM reverse Sequence-Specific Oligonucleotide (SSO) kit. Resolution of the genotype to allele level was performed with the Dynal Biotech sequence-specific priming kits. Any alleles which could not be determined were typed using bespoke sequence-specific primers. All alleles in the international immunogenetics information system (IMGT) library (version 2.4.0) were considered (Robinson *et al.*, 2014).

3.4 Participant selection for this study

First, suitable participants had to be chosen for this study. This study only compared factors linked to untreated HIV disease progression. Therefore, all samples from both the experimental and test groups were HIV positive. To ensure that our collaborators had enough sample left over for each participant for the primary aims of the larger study, sample selection criteria as specified by them, in addition to that required for the metabolomics investigation as specified by us, was considered. Figure 3-2 shows an overview of the selection of participants.

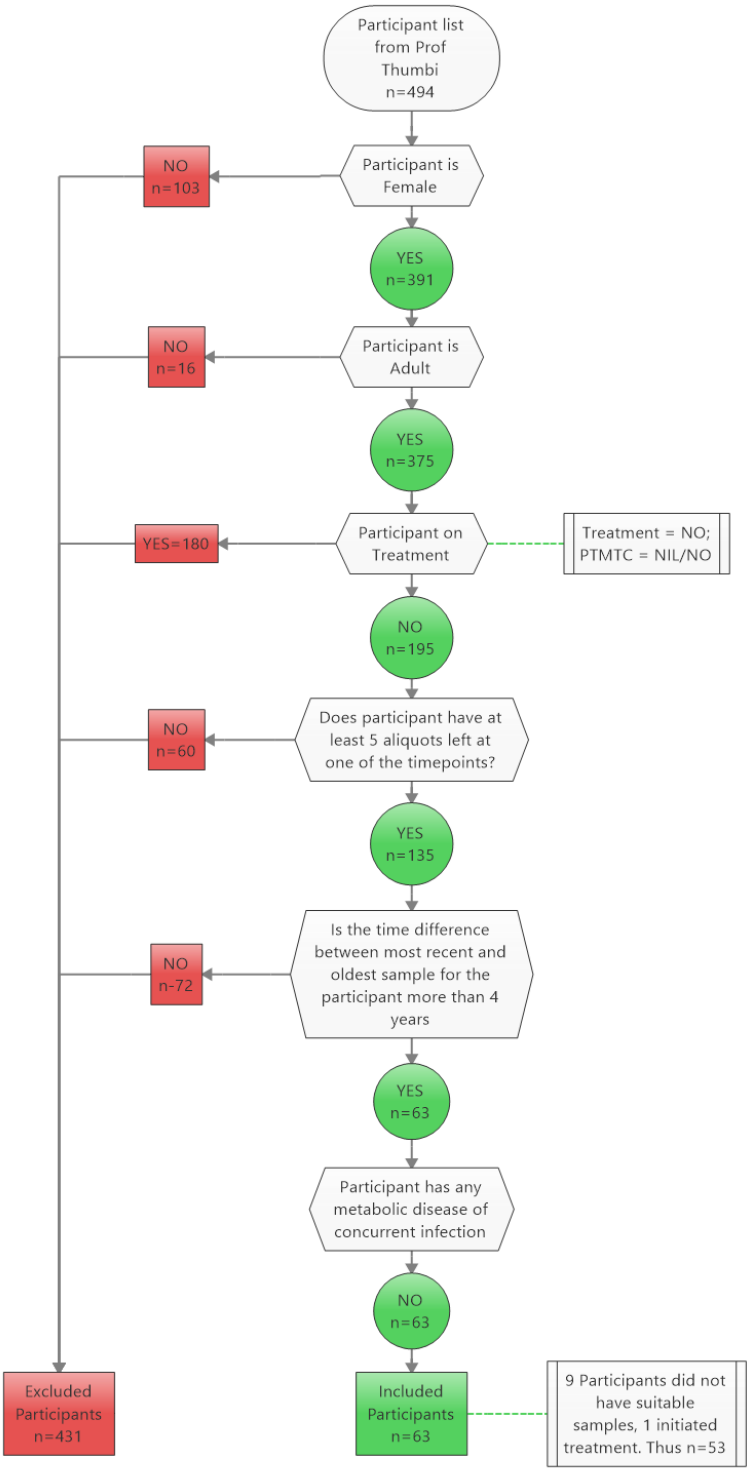
Participant list from Prof Thumbi n=494
Participant is Female
NO n=103
YES n=391
Participant is Adult
NO n=16
YES n=375
Participant on Treatment
YES=180
Treatment = NO; PTMTC = NIL/NO
NO n=195
Does participant have at least 5 aliquots left at one of the timepoints?
NO n=60
YES n=135
Is the time difference between most recent and oldest sample for the participant more than 4 years
NO n-72
YES n=63
Participant has any metabolic disease of concurrent infection
NO n=63
Excluded Participants n=431
Included Participants n=63
9 Participants did not have suitable samples, 1 initiated treatment. Thus n=53

Figure 3-2: **Flowchart showing the selection of participants. Cases were included in the study if they were HIV positive, adult, females for whom there was sufficient sample aliquots and longitudinal data available. Cases were excluded if pregnant, on anti-HIV treatment, experiencing opportunistic infections and/or presenting with metabolic disease. Working through these filters and criteria reduced our cohort size from 494 to 53 participants.**

The demographic and clinical data of the study participants from which sample selection could be made was considered for further selection. This data allowed for filtering of cases from n=494 to n=53 as per the criteria specified in Table 3-1. For homogeneity of the metabolomics cohort, it was decided to use samples collected from only HIV+, non-pregnant female participants. Of the 494 HIV-infected population assigned to us, 391 were non-pregnant adult females, and 103 were males. The 391 selected female participants were filtered for only adults, eliminating an additional 16 children. Treatment would serve as a confounder of the natural control of HIV infection, and hence, any participants on treatment (n=180) were excluded. The remaining 195 participants were subsequently filtered for clinical data and sample availability. One hundred and thirty-five of these participants had more than 5 aliquots available, at a single time point. Given that the time factor associated with determining the rate of progression is a minimum of 4 years, only samples from 63 of these participants were chosen. In addition, criteria were set to exclude HIV positive individuals with concurrent opportunistic and/or viral infections, or those with previously identified metabolic diseases, since these parameters may interfere with the metabolic profiles. However, no such cases were identified in this study. The criteria set out in Table 3-1 aims to homogenise the sample cohort concerning descriptive and clinical criteria, thereby ruling them out as confounders to those factors influencing HIV disease progression. Using the criteria summarised in Table 3-1, and applying this as stated in Figure 3-2, the 63 participants deemed suitable for this study were selected. Nine of these participants, however, were later excluded for not having samples within the specified criteria and another for initiating treatment (explained in the next section).

Table 3-1: Table showing the criteria specified by collaborators and/or as required for metabolomics investigation

Inclusion criteria	Exclusion criteria	Justification
HIV-infected		The aim of this study revolves around HIV disease progression; hence, the test and control groups are HIV positive.
Female		More than ¾ of the participants in the initial study were female. Females provide a homogenous group for metabolomics analysis. Females are also generally reported to be slow/non-progressors (Siddiqui *et al.*, 2009).
Adult		The majority of participants were adult. Metabolic variation in children is known to be age dependant (Gu *et al.*, 2009).
	On treatment	Treatment metabolites, as well as secondary effects, will increase the cohort variation.
	Pregnant	Pregnancy can alter the metabolic profile.
Have more than 5 aliquots left at one time-point		Prerequisite from UKZN.
Participated in the study for at least 4 years		To determine clinical progression, participants had to be followed up for a minimum of at least 4 years.
	Concurrent opportunistic and/or viral infections	Concurrent infections can influence the immune system and the metabolic profile.
	Known metabolic disease	Metabolic diseases have significantly altered metabolite profiles when compared to individuals without metabolic diseases.

3.5 Time point/sample selection

After filtering participants according to the criteria set out in Section 3.4, 63 participants were deemed suitable for the metabolomics investigation of untreated HIV disease progression. The next step was to select which samples (time points) from these 63 participants to use to make up the longitudinal CD4 and *HLA-B* genotype cohorts.

3.5.1 Scope of time point/sample selection

Since our collaborator agreed to share 100 samples and only 63 participants met the inclusion and exclusion criteria, it was decided to include more than one sample per participant representing time points with different CD4 counts. This was decided so as to investigate the plasma metabolic profiles of untreated HIV positive participants with (1) different CD4 counts, (2) median CD4 counts but presenting with slow and fast declines in CD4 count over a time period of more than 4 years, respectively and (3) median CD4 counts but presenting protective vs. non-protective *HLA-B* alleles, respectively. Time point selection was crucial to minimise variation caused by CD4 count in the sub-studies that do not focus on comparing samples at different CD4 counts.

Additionally, to the CD4 counts, stage of HIV disease progression could also possibly cause metabolic variation. Unfortunately, clinical staging of participants could not be determined as the patients were already infected for an unknown period time before recruitment in the primary study at UKZN.

3.5.2 Selection of samples (time points) based on CD4 count

Initially, the untreated HIV patients' plasma samples were stratified into three groups according to their CD4 values being either: 1. high (500 cells/µl <CD4<800 cells/µl), 2. mid (350 cells/µl <CD4<499 cells/µl) and 3. low (CD4<250 cells/µl). The first sub-study investigated the varying metabolic profiles of plasma samples collected from untreated HIV individuals with different CD4 counts. Therefore, a comparison of the plasma metabolome of the high- and low-CD4 groups was made. For the second and third sub-studies, only untreated HIV positive plasma samples in the mid-CD4 range was used, in order to eliminate the possible confounding effects of CD4, when investigating the altered metabolome associated with HIV disease progression. The average CD4 counts were calculated for each of the respective groups: low=215 cells/µl, mid=422 cells/µl, and high=592 cells/µl. Next, from the available sample collection time points, the time point with the closest CD4 count to the average of each CD4-group for each patient was determined.

For each participant, 3 time points closest to the CD4 count of the low, mid and high groups were selected. Even though the time point with the closest CD4 count to the average was selected, that time point's CD4 count didn't necessarily fall into that group's range (i.e. low: CD4<250 cells/µl). The samples that satisfied the specific group's CD4 count criteria were chosen as the final samples for this study. Nine participants were excluded for not having any samples within the criteria, and another was excluded after the participant was determined to be on treatment.

Therefore, the final number of participants used in this study was 53, and the final number of samples used was 96.

Since samples from the low, mid or high groups were selected from each patient if they satisfied the criteria, some participants had samples in all three groups while others only had samples in a single group. Figure 3-3 shows the distribution of participants between the three CD4 count groups.

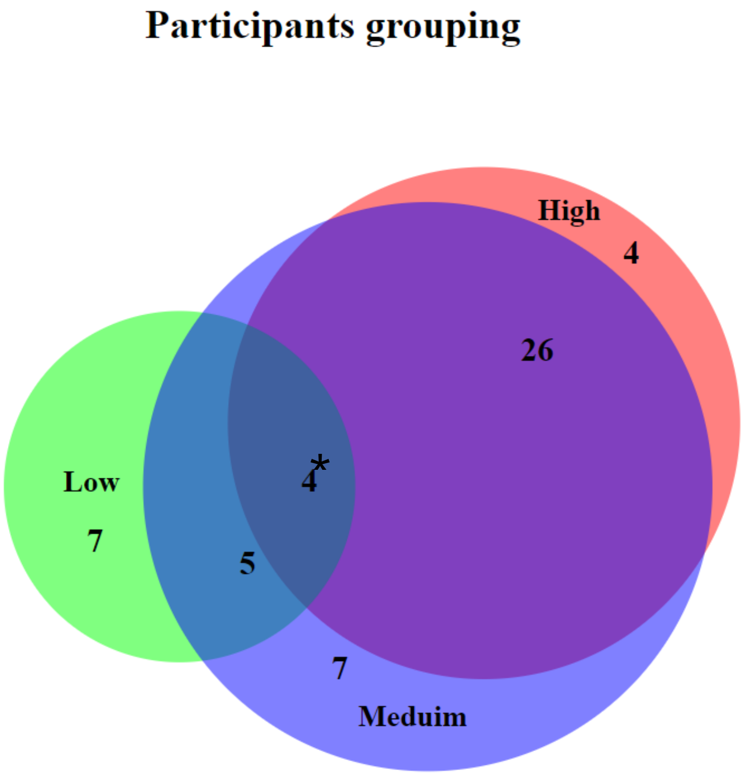

Figure 3-3: Venn diagram showing the distribution of the 53 participants' samples in the defined CD4 groups.

3.6 Participant samples stratified according to CD4 count

The CD4 count and viral load are the preferred parameters for monitoring HIV disease progression. Previously, anti-HIV treatment was only issued to individuals with CD4 counts below 350 cells/µl blood. Fortunately, these globally applied guidelines were changed in September 2016, and all HIV-positive individuals are now eligible for anti-HIV treatment, irrespective of CD4 counts. Generally, regarding HIV prognosis, lower CD4 counts are associated with more severe disease, whereas higher CD4 counts are associated with less severe disease. By comparing the metabolite profile of high and low CD4 counts, insights can be gained as to the role of CD4 as a parameter to assess HIV disease progression.

Section 3.5 outlines the details of the sample selection. Four participants had samples in both the high and low CD4 groups. Since the statistical tests are independent, each participant could only have one sample. Given that the CD4-high group already had a sufficient number of samples, the four samples which overlapped in the low and high CD4 groups (marked by an asterisk in Figure 3-3), were assigned to the CD4 low group. The samples used for this sub-study were thus CD4-high (n=30) and CD4-low (n=16).

3.7 Participant samples stratified according to CD4 change over time

An HIV positive patient's CD4 counts are, unfortunately not the only determinant of HIV progression. Various other factors including age, time of day, other infections etc. also influence these CD4 counts. Within-subject variation for CD4 counts has also been reported to have a coefficient of variation (CV) of up to 25% (Hughes *et al.*, 1994). Considering this, it becomes clear that although interpreting a patient's CD4 counts alone has prognostic value, the change in CD4 count over time ultimately defines disease progression (Taylor *et al.*, 1989). The change in CD4 over time is expressed as a change in cells/µl per time interval. It is calculated from a linear regression model applied to clinical readings of CD4 counts over time. Some studies determine the aforementioned slope using as few as 2 or 3 points (2 to 3 years) (Mellors *et al.*, 2007) while others state that a minimum of at least 8 years' worth of CD4 counts are necessary for accurately determining HIV disease progression (Grabar *et al.*, 2009; Olson *et al.*, 2014). Considering the 25% CV in CD4 measurements in HIV positive individuals, as observed by Hughes *et al.* (1994), studies of HIV disease progression utilising only 2 or 3 measurements are error-prone. For this reason, it was decided to use CD4 count data over a minimum of 4 years for assessing progression.

In HIV disease progression, VL increases with the decrease in CD4 count. For this reason, many studies add viral load as an additional parameter to assess HIV disease progression. Our cohort, unfortunately, had many time points where VL was not measured. It was therefore decided to define our respective cohorts using the CD4 parameter [low vs high (sub-study 1) and median CD4 (sub-study 2 and 3)].

The slope of CD4 change is generally reported in the literature as the defining parameter of HIV disease progression. The equation for slope does not take variation into account (see equation below). Therefore, it is generally used in conjunction with the r-squared value (evaluation of linear model fitting) and p-value (probability). R-squared values closer to 1 indicates a good fit, while lower values indicate a bad fit. Small p-values, on the other hand, indicate better prediction and higher p-values show less accurate prediction accuracy. The equation for the slope based on linear regression is as follows:

$$slope = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

X and y represent the variables (in this case days since the first visit and the absolute CD4 count, respectively) and n the number of observations (time points).

Another model for assessing the relationship between two variables is a correlation. The significant difference is that the equation for correlation coefficient considers standard deviation. The correlation coefficient is a value between -1 and 1. Correlation coefficients closer to -1 or 1 indicate good negative association and good positive correlation, respectively, while values closer to 0 indicates a bad association. Since the correlation coefficient includes standard deviation, it inherently represents fitting, and the additional r-squared value is not required. Correlation coefficients greater than ±0.5 are considered significant (Sullivan & Feinn, 2012). The equation for the correlation between two variables is as follows:

$$Correlation\ coeficient(r) = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{nSD_x SD_x}$$

X and y represent the parameters measured (in this case days since the first visit and absolute CD4 count, respectively), n the number of observations (time points), SD_x the standard deviation in x and SD_y the standard deviation in y.

Figure 3-4 shows the relationship between the slope (x-axis) and the correlation coefficient (y-axis) of all study participants. All participants with a significantly negative correlation coefficient comparing CD4 counts and elapsed time (≤-0.05) had a slope of fewer than -0.03 cells/µl per day. However, when participants with a slope of fewer than -0.05 cells/µl per day are considered, the worst correlation coefficient achieved was -0.13, which is insignificant. Since there are statistical cut-offs for the significance of correlation coefficients, it was decided to use correlation coefficients as an indicator for HIV disease progression, rather than slope, to group the participants as progressors or non-/slow-progressors based on CD4 change over time. Participants were defined as progressors (n=21) when presenting with a significant negative correlation between CD4 counts and time (≤-0.05) while all other participants were defined as non-/slow-progressors (n=21).

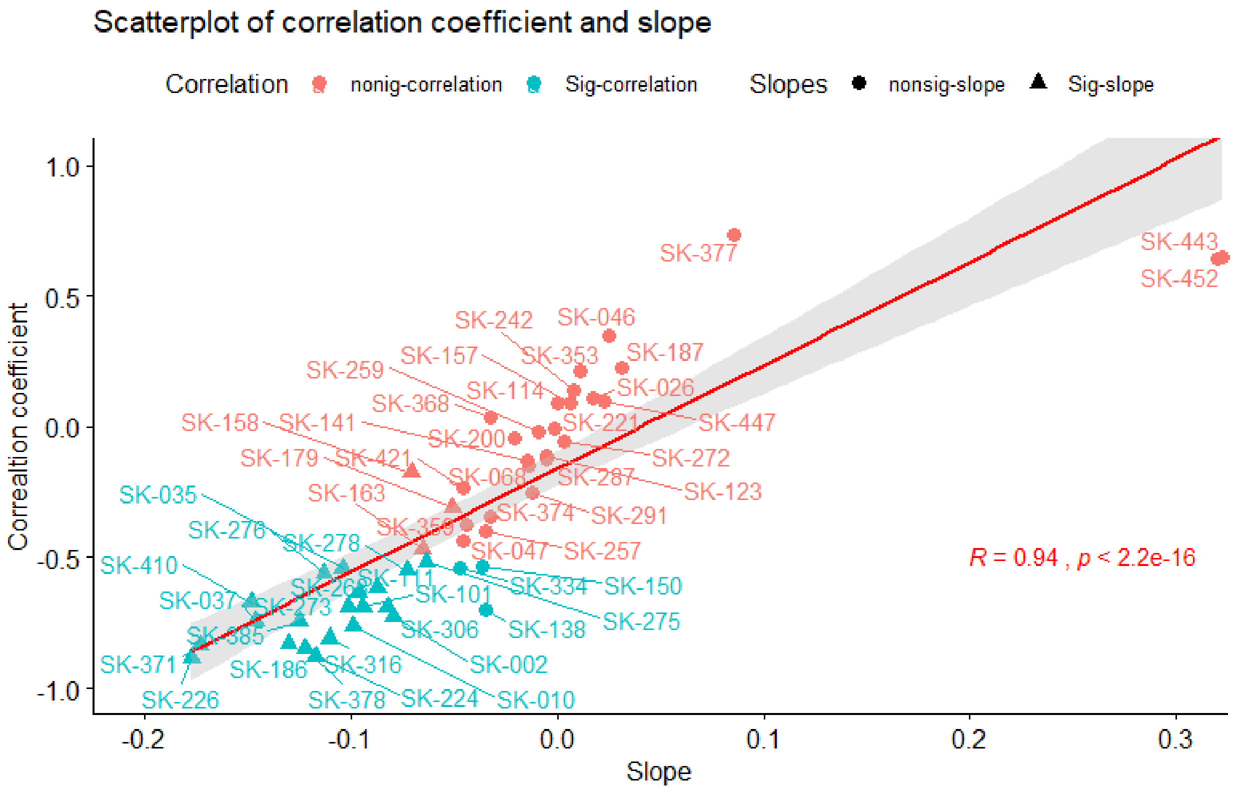

Figure 3-4: **Scatterplot of the correlation coefficients and slopes of CD4 counts over time for all participants with a regression line and confidence interval. Correlation coefficient coloured according to significance; non-significant correlation coloured peach and significant correlation coloured blue. Slope points were shaped according to significance; non-significant slope shaped circular and significant slope shaped triangularly. The R-squared and p-values for the regression line are 0.94 and < 2.2e-16, respectively.**

3.8 Participant samples stratified according to *HLA-B* alleles

The third sub-study determined the effect of varying *HLA-B* alleles (a genetic parameter associated with control of HIV infection) on the plasma metabolome, (Carrington & Walker, 2012; Limou & Zagury, 2013). Section 2.4 reviewed the role of HLA in HIV disease progression. Therefore, this section will only discuss the classification of participants based hereon.

Many *HLA* alleles have been identified in previous studies to be predictive of HIV disease progression (Bardeskar & Mania-Pramanik, 2016; Borghans *et al.*, 2007; Carlson *et al.*, 2012; Costello *et al.*, 1999; Fernandes-Cardoso *et al.*, 2016; Flores-Villanueva *et al.*, 2001; Gentle *et* *al.*, 2017; Jiang *et al.*, 2013; Kløverpris *et al.*, 2012; Leslie *et al.*, 2010; Limou & Zagury, 2013; Loubser *et al.*, 2017; Matthews *et al.*, 2011; N. *et al.*, 2014; Payne *et al.*, 2014; Rousseau *et al.*, 2008; Silva *et al.*, 2010; Wright *et al.*, 2010). In our investigation, we stratified the untreated HIV positive participants according to protective and non-protective *HLA-B* alleles. The characterisation of these alleles was previously described by Wright *et al.* (2010), who investigated the underlying mechanisms of HIV control (viral fitness and HIV pathogenesis) as afforded by protective *HLA* alleles. The *HLA-B* alleles identified by the authors to be protective were *HLA-B**57, *HLA-B**58:01 and *HLA-B**81. It was assumed that the protective allele would be dominant, and therefore any participant who is either homozygous or heterozygous for any of these *HLA-B* alleles were classified as having a protective allele. All participants who did not have any of these protective *HLA-B* alleles were, for this study, classified as having non-protective *HLA-B* alleles. Therefore, in our investigation, study participants for accomplishing sub-study 3, were grouped as protective (n=12) and non-protective (n=30), based on the possession of a protective *HLA-B* allele as specified by Wright *et al.* (2010).

3.9 Demographics and clinical data

Discussing participant demographics and clinical data concisely in the context of a complex experimental design such as this is challenging since the demographics and clinical data for sub-study 1 differs from that of sub-study 2 and 3 due to the use of different samples, from the same participants for the respective sub-studies. Hence, participant demographics and clinical data is tabled for the sample group as a whole as well as per individual sub-study.

3.9.1 Demographics and clinical data of the cohort as a whole

Considering the aim, the low CD4 count, progressor and non-protective *HLA-B* allele classifications suggest a "sicker" cohort while the high CD4 count, non-progressor and protective *HLA-B* allele classifications suggest a "healthier" cohort (Figure 3-5). When considering the 53 participants, 33 were classified as "sicker" in some of the sub-studies and "healthier" in others while only 8 and 12 were exclusively "healthier" and "sicker" throughout all three sub-studies. This already shows the disjointedness between the classification of HIV disease progression as dictated by the available sample data and the definition for these as proposed by previous literature.

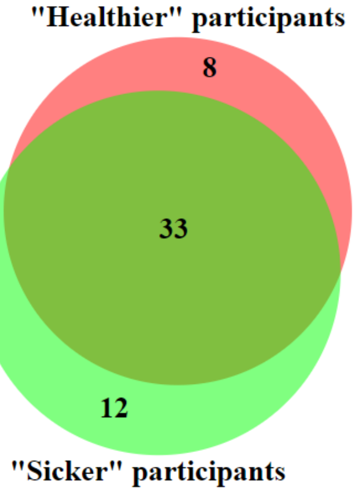

Figure 3-5: Venn diagram indicating the number of participants classified by the different sub-studies as “healthier” or “sicker”.

Demographic and clinical data for the entire cohort is useful for confirming the correct application of the selection criteria. Participants had to have at least one time-point with more than five aliquots of the sample left. Table 3-2 shows that the participant with the lowest number of time points, with at least five aliquots, had three time points with more than or equal to 5 aliquots remaining. To determine the change in CD4 over time, CD4 count data had to be available for at least four years, and therefore the duration of participation. The minimum hereof can also be seen in Table 3-2 to be 4.34 years. The average age of participants was 32 years with a standard deviation of 8.23 years. The youngest and oldest participants were 21 years old and 54 years old, respectively. Considering all time points of the selected participants, the average CD4 count is 444.98 cells/µl blood with a standard deviation of 137.3. The minimum and maximum CD4 count recorded for the 53 participants was 161.53 cells/µl blood and 856.2 cells/µl blood, respectively. The maximum CD4 count reported here is beyond the range specified in the sample selection criteria because all time points were considered here in calculating the statistics shown in Table 3-2.

Table 3-2: Table showing the demographics and clinical data of the entire cohort

Number of participants = 53	Average	min	max	Standard deviation
Number of time points useable	11.98	3.00	21.00	3.98
Duration of participation (years)	6.35	4.34	8.78	1.00
Age at Enrolment (years)	32.86	21.65	54.78	8.23
Average CD4 for duration of participation (cells/µl)	444.98	161.53	856.20	137.30
Change in CD4 over time (cells/µl/day)	-0.04	-0.32	0.32	0.10
Average Viral load (RNA copies/µl)	54002.31	0.00	286421.82	67403.09
Change in viral load over time (RNA copies per day)	1.39	-85.64	137.29	32.12

3.9.2 Participant demographics and clinical data stratified according to high and low CD4 counts.

Table 3-3 shows the distribution of demographic and clinical data between the high-CD4 and low-CD4 groups as used in sub-study 1. No significant differences are evident between the demographic and clinical parameters and the stratification used here except for CD4 and average CD4, as highlighted in Table 3-3.

Table 3-3: Table showing participant demographics and clinical data stratified according to high and low CD4 counts

	High-CD4 n=30				Low-CD4 n =16				p-value
	Average	Min	max	Standard deviation	Average	Min	max	Standard deviation	
Duration of participation (years)	6.4	4.6	8.8	1.0	6.1	4.3	7.9	0.9	0.47178
Age at Enrolment (years)	32.8	21.6	52.8	8.7	31.2	22.5	40.1	5.7	0.52874
Sample age (years)	10.5	7.0	14.4	1.8	9.5	6.8	13.1	1.7	0.063524
Sample CD4 (cells/µl)	585.4	509.0	711.0	47.1	217.3	202.0	240.0	9.3	4.40E-31
average CD4 for duration of participation (cells/µl)	529.1	372.2	856.2	112.0	309.4	161.5	437.0	64.3	8.65E-09
Change in CD4 over time (cells/µl/day)	-0.1	-0.3	0.3	0.1	0.0	-0.1	0.0	0.0	0.69474
Sample VL(N/A at some tps) (RNA copies/µl)	42553.8	2970.0	187651.0	61384.3	149651.7	3161.0	654608.0	229223.6	0.17055
Average Viral load (RNA copies)	43569.9	0.0	242100.7	61680.4	76938.5	15469.0	286421.8	79569.2	0.13061
Change in viral load over time (RNA copies per day)	0.7	-85.6	70.0	25.3	0.0	-42.8	84.3	30.2	0.93051

Since the samples' CD4 counts are the conditional variables in this analysis, it is expected to differ significantly between the groups. The extent of this difference is shown in Figure 3-6. The average CD4 counts over the minimum period of 4 years, was also significantly different between the two groups, and correlated with the actual sample CD4 count of a patient at a specific time point, with a correlation coefficient of 0.77338 ($p<0.001$) as seen in Figure 3-7. This indicates that the samples in the high-CD4 group had higher overall CD4 count values than those in the low-CD4 count group. This difference in average CD4 counts can be seen in Figure 3-8. Although a direct link between time point (sample) CD4 count and average CD4 counts exists, the correlation between them indicates that metabolomics findings are not only applicable to the CD4 count of a patient at the time the sample was taken but also representative of the average CD4 measured over time. Similarly, viral load at a time point (sample) correlated with average viral load, with a correlation coefficient of 0.5637 ($p<0.001$). Viral load and change in viral load correlated with a correlation coefficient of 0.59436 ($p<0.001$). As expected in the "sicker" / CD4-low group, the VL is higher. The relatively high standard deviation (SD) in the groups attributes to why the p-value was not significant.

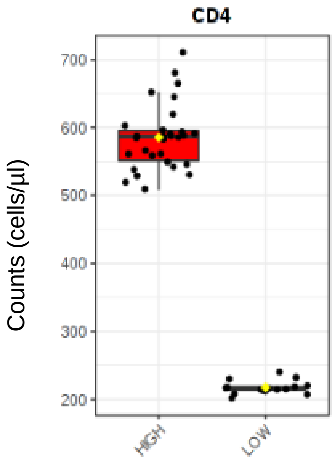

Figure 3-6: Boxplots of the CD4 counts of samples in the CD4-high and CD4-low groups

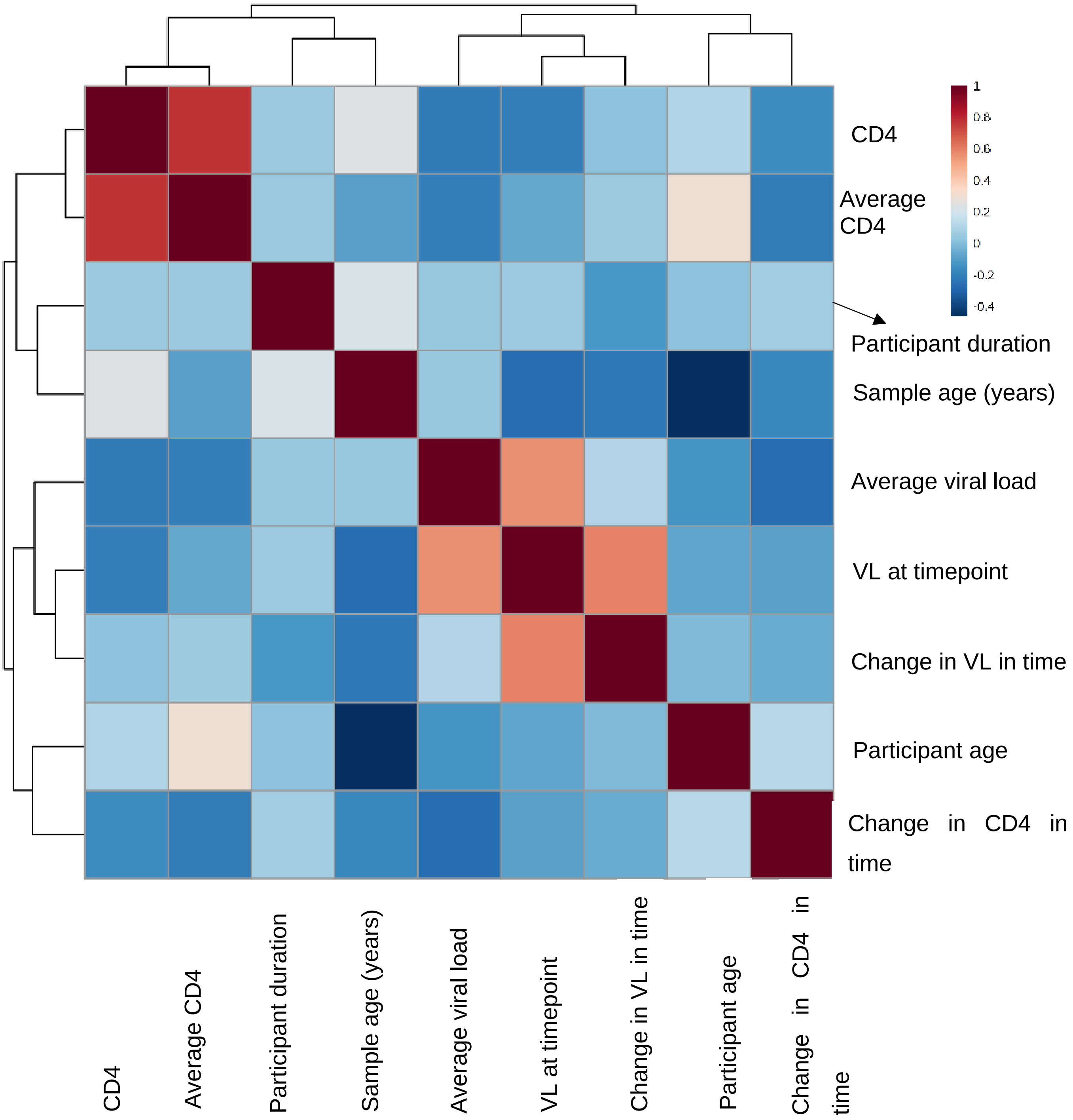

Figure 3-7: Correlations between the demographic and clinical data of all participants. Correlation coefficients are coloured according to the scale.

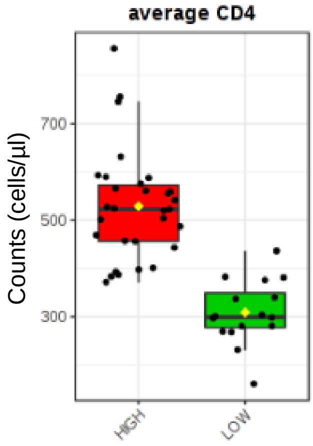

Figure 3-8: Boxplots of the average CD4 count in the CD4-high and CD4-low groups

3.9.3 Participant demographics and clinical data stratified according to the change in CD4 count over time

Table 3-4 summarises the demographic and clinical data of the non/slow-progressors and progressors based on a non-significant correlation with time and a significant negative correlation with time, respectively. The only significant differences between the groups are the change in CD4 counts over time, change in viral load over time and CD4 counts correlation. CD4 counts correlation and change in CD4 counts over time, are expected to confound each other as these involve different approaches for measuring the change in the CD4 counts, using either correlation or slope, respectively. The literature on monitoring HIV disease progression uses CD4 counts and VL load as parameters for such, with CD4 counts decreasing and VL increasing over time, indicating an inverse correlation. This could justify the significant difference seen between non/slow-progressors and progressors, however, the correlation coefficient between these two parameters in this dataset, is only -0.11 (p-value = 0.45) which is not significant. Change in viral load observed over time did not correlate significantly with any of the measured demographic or clinical data parameters. This is indicative of an unaccounted-for parameter driving the change in VL observed over time. For this reason, this parameter can be considered a confounder in this metabolomics analysis. Viral load was higher in the progressor group as expected, but not significantly so, due to the high standard deviation (SD).

Table 3-4: Table showing participant demographics and clinical data stratified according to the correlation between CD4 count and time

	Non/Slow-progressors (non-significant correlation with time) n=21				Progressors (significant negative correlation with time) n=21				
	Average	Min	max	Standard deviation	Average	Min	Max	Standard deviation	p-value
Duration of participation (years)	6.3	5.0	8.7	1.0	6.4	4.6	8.8	0.9	0.84389
Age at Enrolment (years)	34.4	21.6	54.8	9.8	31.0	22.5	42.7	5.6	0.18928
Sample age (years)	10.2	7.2	13.2	1.8	9.9	6.4	13.4	1.9	0.60577
average CD4 for duration of participation (cells/µl)	431.9	298.4	588.1	93.1	473.2	297.3	756.1	107.6	0.19686
Change in CD4 over time (cells/µl/day)	0.0	-0.1	0.3	0.1	-0.1	-0.3	0.0	0.1	4.73E-07
Sample Viral load (N/A at some time points)	82572.1	6240.0	287000.0	102427.3	176167.4	7342.0	854000.0	286060.2	0.576
Average Viral load (RNA copies)	39025.4	0.0	242100.7	52424.3	68198.2	78.3	286421.8	77291.1	0.16671
Change in viral load over time (RNA copies per day)	-9.3	-85.6	27.6	25.2	16.8	-26.7	137.3	36.5	0.011161
Sample CD4 (cells/µl)	413.1	356.0	474.0	24.7	422.1	360.0	469.0	23.8	0.24165
CD4 correlation	0.0	-0.4	0.6	0.3	-0.7	-0.9	-0.5	0.1	8.53E-13

3.9.4 Participant demographics and clinical data stratified according to protective and non-protective *HLA-B* alleles

Table 3-5 shows the demographics and clinical data of participants with protective and non-protective *HLA-B* alleles. The *HLA-B* genotypes op the participants are summarised in Table 3-6. There were no significant differences in the demographic and clinical data parameters displayed here between the two groups. Since the change in CD4 counts and VL in time was used in previous literature to characterise *HLA-B* alleles as protective or non-protective, a significant difference in these parameters is expected. For this specific cohort, participant samples defined as having protective and non-protective *HLA* alleles did not present with a non/slow-progressor and progressor phenotypes, respectively as based on the clinical parameters CD4 counts and VL. Considering the distribution of the *HLA-B* alleles (Table 3.6), it is evident, using the definition of Wright et al that 28% of the alleles were of the protective phenotype whilst 72% were of the non-protective phenotype.

With reference to HIV, not all the *HLA-B* alleles have been characterised hence their role with regard to protection and/or non-protection is unclear for many alleles. Our strict classification as per Wright et al i.e. that *HLA*-B*57, *HLA-B**58:01 and *HLA*-B*81 alleles are protective, does not rule out the fact that the 2nd allele may be a deleterious one. It was assumed that the protective allele is homozygous or heterozygous dominant. The presence of a possible deleterious allele would therefore impact on the functionality of HLA and the associated immune and metabolic profiles measured. Since assumed to be recessive here, the deleterious' allele's impact may be negligible. Since the role of many *HLA-B* alleles remains undefined, this may also be a contributing factor to the overlapping PCA profiles. Future work may entail further subdivision of the *HLA-B* alleles into protective and non-protective and characterising the metabolic profiles of these.

Table 3-5: Table showing participant demographics and clinical data stratified according to protective and non-protective *HLA-B* alleles

	Protective n=12				Non-Protective n=30				
	Average	Min	Max	Standard deviation	Average	Min	Max	Standard deviation	p-value
Duration of participation (years)	6.24	4.61	8.74	0.94	6.65	5.31	8.78	0.99	0.23
Age at Enrolment (years)	33.25	21.65	54.78	8.84	31.11	23.21	42.70	6.03	0.46
Sample age (years)	10.07	6.40	13.37	1.76	10.05	7.24	13.20	2.07	0.98
average CD4 for duration of participation (cells/µl)	446.24	297.33	593.26	93.83	444.68	317.88	566.35	89.94	0.96
Change in CD4 over time (cells/µl/day)	-0.04	-0.18	0.32	0.09	-0.05	-0.12	0.03	0.05	0.87
Viral load of sample (N/A for some samples)	146996.10	6240.00	854000.00	159977.72	84757.60	N/A	287000.00	78398.16	0.70
Average Viral load (RNA copies)	51093.35	78.33	286421.82	67442.90	58385.84	2767.70	242100.71	67068.54	0.88
Change in viral load over time (RNA copies per day)	4.30	-85.64	84.27	24.33	6.05	-85.64	137.29	51.06	1.88
Sample CD4 (cells/µl)	421.53	356.00	474.00	24.90	409.00	356.00	469.00	35.50	0.15
CD4 correlation	-0.32	-0.89	0.63	0.41	-0.40	-0.88	0.25	0.36	1.15

Table 3-6: **Summary of *HLA-B* alleles of all the participants. Alleles marked with an asterisk (*) were used in the stratification of participant samples as having protective alleles as per the definition of Wright et al 2010. All other alleles were regarded to be non-protective.**

HLA-B **Allele**	**Protective**	**Non-protective**	**Total**
07:00	0	4	4
07:02	1	0	1
08:00	0	3	3
08:01	1	1	2
13:00	0	1	1
13:02	1	0	1
14:01	2	0	2
14:02	0	2	2
15:00	0	0	0
15:03	1	6	7
15:10	0	6	6
15:16	0	2	2
18:00	0	3	3
35:00	0	2	2
39:00	0	1	1
39:10	1	2	3
40:01	0	1	1
40:06	0	1	1
42:01	1	9	10
42:02	1	0	1
44:00	0	7	7
44:03	0	1	1
45:01	0	3	3
51:00	0	1	1
53:01	0	1	1
57:02*	1	0	1
57:03*	1	0	1
58:01*	5	0	5
58:02	0	2	2
67:00	1	0	1
81:00*	1	0	1
81:01	4	0	4
82:02	2	0	2
15:03/29	0	1	1
Frequency	29%	71%	

METHODS AND REPEATABILITY

This chapter includes the methods used for the analysis of samples as well as results pertaining to the standardisation of the methods to assess their performance on the grounds of repeatability.

4.1 Chemicals

For this metabolomics study, several chemicals were used during the extraction and derivatisation process. All water used was acquired from the Millipore Elix Advantage and Milli-Q Integral water purification system for ultrapure water. Acetonitrile, chloroform and methanol were purchased from Honeywell (Charlotte, North Carolina, USA). 3-phenyl butyric acid, norleucine, nonadecanoic acid, acetaminophen, pyridine, methoxyamine hydrochloride and n,o-bis(trimethylsilyl)tri-fluoroacetamide [BSTFA] [with 1% trimethylchlorosilane (TMCS)] was purchased from Sigma-Aldrich (St. Louis, Missouri, USA). Nitrogen and Helium baseline 5.0 was purchased from African Oxygen Limited [AFROX] (Johannesburg, South Africa).

4.2 Equipment and consumables

Several preparatory and analytical equipment, as well as consumables, were used to prepare the plasma samples for analysis. Eppendorf 2 ml DNA LoBind tubes, Agilent 2 ml glass vials and caps and Agilent 250 µl Glass pulled-point conical inserts were purchased from Sigma-Aldrich (St. Louis, Missouri, USA).

The Leco Pegasus 4D analytical system was used for the analysis of plasma samples. This system was previously purchased from Leco (St. Joseph, Michigan, USA).

4.3 Reagent and sample preparation

4.3.1 Preparation of reagents

The internal standard and the methoxyamine solutions were prepared once and used throughout the study. This paragraph explains the preparation of each.

4.3.1.1 Internal standard (IS)

Thirty-six milligrams of each of the internal standard (3-phenyl butyric acid [3-PHB], norleucine, nonadecanoic acid and acetaminophen) was weighed off in separate glass weighing boats. The weighed standards were washed off into a 100 ml volumetric flask using 20 ml chloroform (nonadecanoic acid), 60 ml methanol (norleucine, acetaminophen and 3-phenyl butyric acid), and 20 ml water to produce the stock IS solution with a final concentration of 360 µg.ml^{-1}. The solution was sonicated for 20 minutes to ensure that internal standards have dissolved. The stock IS was diluted 1:10 using a 1:3:1 solution of chloroform, methanol and water to obtain a final concentration of 36 µg.ml^{-1}.

4.3.1.2 Methoxyamine solution

Methoxyamine hydrochloride (300mg) was weighed in a weighing bottle and transferred to a 20 ml volumetric flask. Any residual methoxyamine hydrochloride was washed into the volumetric flask with pyridine and made up to a volume of 20 ml.

4.3.2 Aliquotting samples and pooled QC-sample preparation

Although ethical approval had been granted for the primary studies where these samples were used, explained in Section 3.3, the collaboration and metabolomics investigation reported here was also approved by the Health Research Ethics Committee (HREC) of the Faculty of Health Sciences at the NWU (reference number NWU-00125-17-A1).

Samples from the UKZN were transported to the NWU on dry ice. The boxes with frozen samples were sequentially removed from the dry ice and left to thaw at room temperature in the laminar flow cabinet. Once thawed, 50 µl of each sample was aliquoted under sterile conditions, into a prelabelled 2ml Eppendorf LoBind tube (minimal leaching of plastics into sample) for later extraction and another 50 µl into a 50 ml tube for the preparation of the pooled QC sample. This was repeated for all samples.

The pooled QC sample allows for method standardisation as it is representative of the entire sample group, thus allowing for monitoring the quality of the analysis. The QC samples are analysed multiple times per batch. If any drift in analytical performance occurs, it will be detectable in the QC samples which can either be used to correct these drifts or mark the batch as unsuitable and due for a re-extraction or rerun. The QC sample is also used to condition the instrument for the samples about to be run (further explained in Section 4.4.2).

Once all samples were aliquoted, the tube containing the pooled QC sample was mixed through several inversions whereafter 50 µl aliquots were transferred into thirty tubes labelled “QC”. The plasma samples and QC aliquots were reorganised into extraction batches, each containing 12 randomised samples and a QC. These samples were then stored at -80°C until the day of the extraction.

4.3.3 Extraction and derivatisation method

Extraction and analysis were performed in batches. This was done to minimise the time samples had to stand on the instrument after derivatisation, as this time is known to influence sample integrity. Each batch was removed from the freezer on the morning of analysis and left to thaw at room temperature for 20 minutes. After thawing, 50 µl Internal standard, dissolved in a 1:3:1 ratio of chloroform:methanol:water, was added to each sample (test & QC) under a fume cabinet. Thereafter, the samples were placed on ice, and 300 µl ice-cold acetonitrile added. The mixture was vortexed for 2 minutes at maximum speed and then incubated on ice for 10 minutes. The insoluble, denatured protein was precipitated through centrifugation at 1413 x g for 10 minutes whereafter the supernatant was transferred to a 2ml Agilent glass vial. The samples were dried under a gentle stream of nitrogen on a heat block set to 30°C for 45 minutes or until dry. Methoxyamine solution (25 µl) as prepared in 4.3.1.2 was added to the dried sample and vortexed for 30 seconds. The vials were capped and incubated on a heat block at 50°C for 90 minutes. After removing the vials and allowing them to cool to room temperature, 40 µl N,O-Bis(trimethylsilyl)trifluoroacetamide (BSTFA) with +1% trimethylchlorosilane (TMCS) was added. The vials were recapped and incubated for another 60 minutes at 60°C. The vials were removed and again allowed to cool to room temperature. The contents were transferred to a pulled point conical insert using a glass transfer pipette. The insert was placed back into the GC vial and recapped. The samples were loaded onto the instrument autosampler, QC first followed by the samples, in the injection sequence explained in 4.4.2

4.4 Instrument methods used during analysis

Before analysis of the samples, the default methods were tested and standardised. The standardisation of GC and MS methods, sample handling on the instrument, standardisation of the data processing method, the alignment method and the normalisation methods are described in this Section.

4.4.1 Standardisation of the GC and MS method

Each project analysed on a sophisticated platform such as the Leco Pegasus 4D requires at least some manual standardisation. The standardisation can vary from method development to slight adjustments of previously developed method parameters. For this project, a standardised extraction method previously optimised for plasma samples was used without modifications (Parihar *et al.*, 2017). The generic analysis methods, on the other hand, could not be used as is, as several factors affect the endpoint parameters to be used. For analysis, the parameters that were standardised included the detector voltage and split ratio. Both of these parameters affect the intensities of the peaks that would be obtained. The split ratio determines the ratio between the injected sample that reaches the column and the sample that was “blown-off” as a result of the split flow. The detector voltage determines the sensitivity of the detector, the more sensitive, the higher the intensities of the peaks, but also the noise.

The split ratio was standardised by running a derivatised QC extract under different GC methods decreasing the split ratio from a fifty split to a three split sequentially while evaluating the chromatogram in-between analyses. The total ion chromatograms were visually inspected for saturation (high peaks with flat or jagged tops). Split ratios between 1:3 and 1:20 yielded several peaks that were saturated. Split ratios between 1:20 and 1:50 did not have any saturated peaks in the QC sample. The standardised ratio of 1:20 was chosen as no saturation occurred while retaining as much trace compounds as possible.

The detector voltage was standardised by running samples with a 1:20 split and adjusting the detector voltage upwards from the tuned detector voltage. The standardised tune detector voltage was 1200V. The highest unsaturated intensities were achieved at a detector voltage of 1400V. This represents a 200V offset from the standardised tune voltage. The MS-method used for further analysis was set at a detector voltage with a constant offset of 200V from the most recent standardised tune voltage, thus running at 1400V.

4.4.2 Autosampler and injection methods

The samples were loaded onto the chilled tray of the autosampler. The first batch was loaded with the QC in position one, followed by the 12 randomised samples. Subsequent batches were loaded in the same order. Once the sequence started, the autosampler performed a wash step, washing the syringe with hexane three times. This wash step was repeated before and after injecting each sample. The autosampler injected 1µl of the sample into the split/splitless inlet once the GC and MS parameters reached their setpoint.

The injection sequence for each batch was as follows:

Qp	Qp	Qp	H	Q1	S1	S2	S3	S4	H	Q2	S5	S6	S7	S8	H	Q3	S9	S10	S11	S12	H	Q4

With Qp representing injections of QC samples to prime the system, H representing an injection of pure hexane to wash any carry over off the column, Q1-4 representing the QC sample injections used to assess the batch quality and possibly perform batch corrections with and S1-12 representing the 12 randomised samples in the batch. Since the QC samples may represent any drift in instrumental factors, these were consistently injected after the column was washed with hexane. These QCs also primed the column for subsequent sample analysis.

4.4.3 Standardised GC method parameters

An Agilent 7890B GC system modified with a modulator and secondary oven was used for GC separation in this study. The injector was operated in the split mode with a Restek Topaz 4.0mm ID low pressure drop inlet liner with wool and maintained at a temperature of 270°C. Helium was used as a carrier gas with a flow rate of 1ml per minute constantly corrected using pressure ramps. A septum purge of 3ml per minute was applied to minimise contamination by or through the septum. A fraction of the 1µl sample injected was split off using a 1:20 split ratio. The GC-column configuration used was a 30m Restek Rxi-5Sil ms column joined to a 1.5m Restek Rxi-17 ms column. The union between the columns was placed as close as possible to the modulator.

Figure 4-1 shows the temperature programming used during the analysis of each sample. The oven temperature was kept at 55°C for 2 minutes whereafter it was heated at 4°C per minute to 300°C for the remainder of the run. The secondary oven was initially kept at 70°C for 2 minutes whereafter it was heated to 300°C at 4.5°C per minute and kept for the remainder of the run. The modulator was initially kept at 85°C for 2 minutes whereafter it was heated to 310°C at 4°C per minute and kept for 12 minutes.

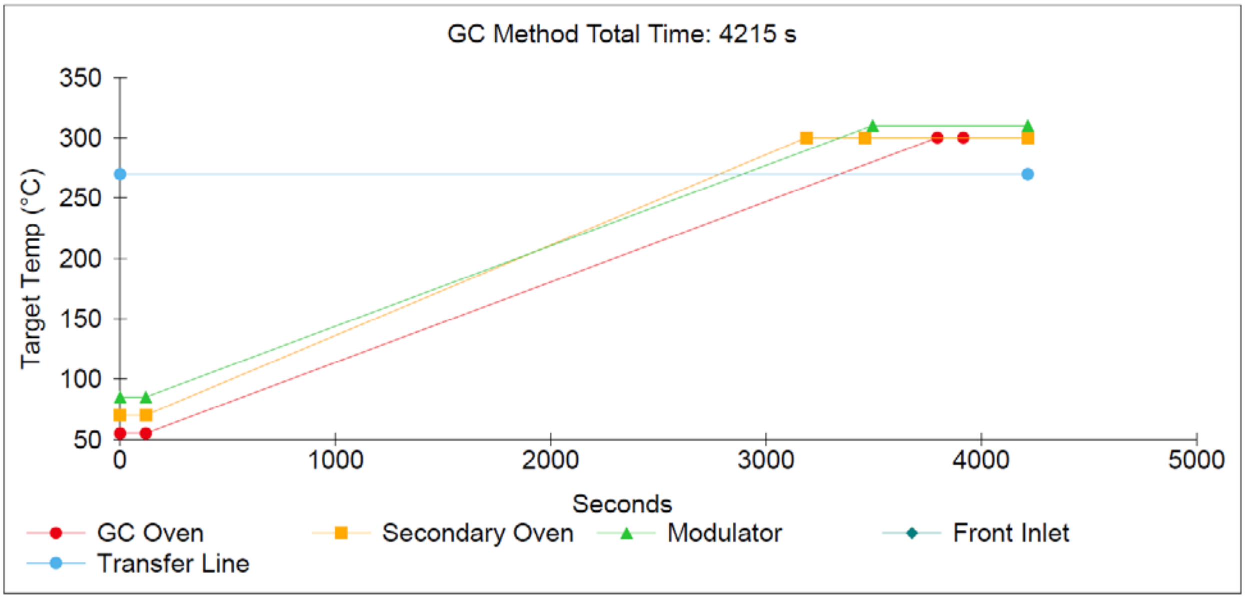

Figure 4-1: GC method temperatures used during the analysis of samples.

The modulation was achieved through 3 second modulation periods with a 0.5-second hot pulse and a 1 second cool between stages. The transfer line was kept at 270°C throughout the entire run.

4.4.4 MS method parameters

Leco's time-of-flight MS was used in this study. The filament was kept off for the first 425 seconds of the run to allow the solvent to elute without being ionised and saturating the detector. After the solvent delay, the masses 50-800 m/z were collected at 200 spectra per second. The tuned detector voltage was increased by 200V hence the final acquisition voltage during standardisation was set to 1400V for analysis. Fragmentation was performed at the standard -70eV. The source temperature was kept at 220°C.

4.4.5 Standardisation of the data processing method

The raw data was processed with the vendor software, Chromatof (Leco, USA). The baseline was calculated, peaks were detected, identified, and the areas thereof calculated. The default data processing method with the following parameters was used as a comparison to the standardised method. The default baseline parameter was set to just above the noise. The default peak width was set at 9 seconds for the first dimension and 0 seconds for the second dimension. The minimum spectral match required to combine modulated peaks was set to 650 (65%). A signal-to-noise ratio of 50 and a minimum of 2 apexing masses was used to separate peaks from the noise. For peak identification in the default method, spectra were searched against the National Institute of Standards and Technology (NIST) 08 library with a minimum similarity match of 800 (80%) before a name was assigned. The unique mass of each peak was then used for area calculations.

The data processing method which yielded the best results was adjusted from the default method with the following parameters. The baseline offset was increased, this reduced variation in area calculation when integrating the peaks. The peak widths were adjusted to 6 seconds for the first dimension and 0.2 seconds for the second dimension. By reducing the first dimension peak width, more peaks were found, especially very small peaks with short retention times. The second dimension width criteria were added to reduce the chances of noise being identified as peaks. The minimum spectral match required to combine peaks in the second dimension was reduced to 600 (60%) to minimise the splitting of peaks and thereby increasing the accuracy of the area calculation. The minimum sub-peak signal-to-noise ratio was decreased to 15 for better coverage and integration over multiple modulations. The signal-to-noise ratio was increased to 100 and the number of apexing masses to 3 to reduce the presence of false peaks. To reduce the number of unknowns, the minimum library match required before a name is assigned was reduced to 750 (75%).

Figure 4-2 shows the QC samples' CV curves of the areas and the normalised concentrations (3-Phenylbutyric acid normalisation explained in Section 4.4.7) of both data processing methods. Since untargeted metabolomics is an unbiased method, a CV cut-off of 50% is generally used as opposed to the 30% untargeted cut-off proposed by the food and drug administration (FDA) (Schoeman & Du Preez, 2012). When considering this CV cut-off, the original method yielded 186 analytes while the standardised processing method yielded 261 analytes (indicated as red lines in Figure 4-2). This equates to 1.4 times more peaks detected in the standardised method with a CV of less than 50%. The standardised method outperformed the original method in respect of these samples.

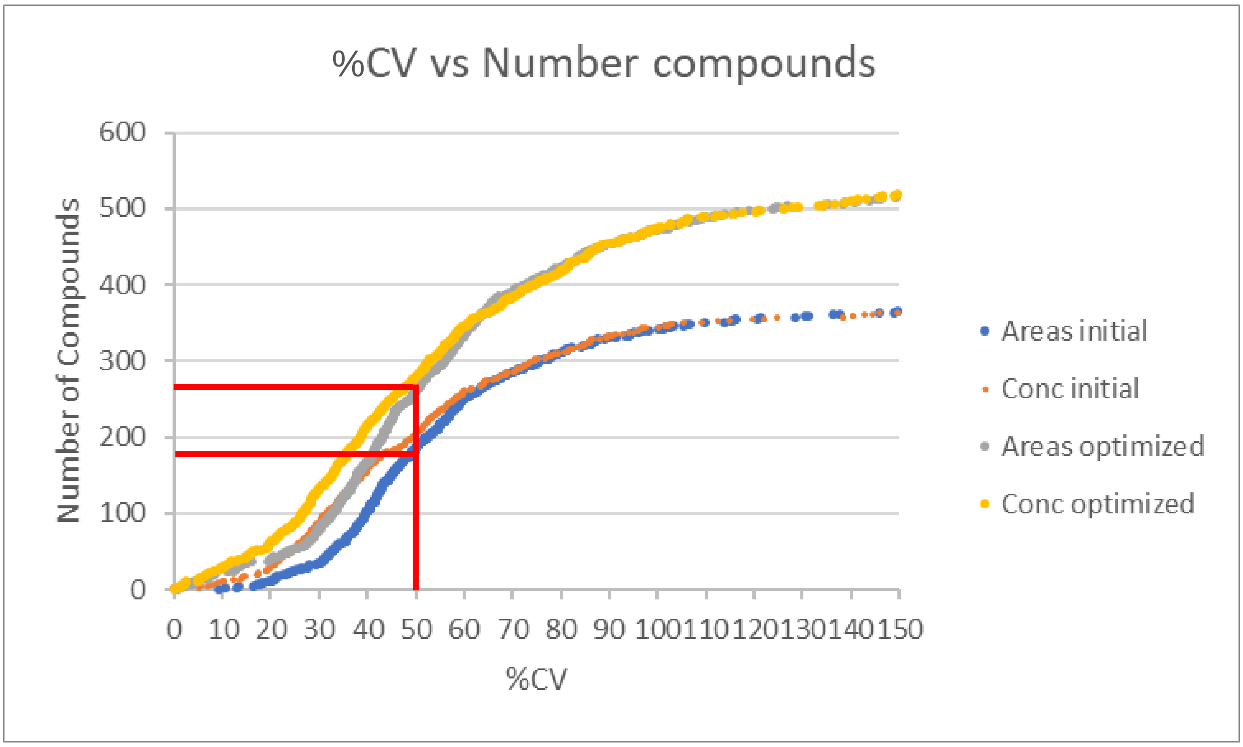

Figure 4-2: CV-curve of initial versus standardised data processing method. The initial method yielded less than 200 compounds with a CV percentage of less than 50. The standardised method yielded more than 250 compounds with CV percentages of less than 50.

4.4.6 Alignment method

For alignment of the peaks from all samples analysed, the spectral match was set with a mass threshold of 20 (2%) and a minimum similarity match of 600 (60%). Retention match was set with the maximum number of modulation periods apart as one and maximum retention time difference as 7. Peaks not found by the initial peak finding parameters were re-evaluated at half the original signal-to-noise ratio. All matches were kept even if an analyte was only detected in a single sample. A total of 210 compounds were aligned and exported. The peak IDs (best spectral match in the library), the unique mass and the two retention times were appended in the general format of PeakID_Uniquemass_RT1_RT2 to ensure uniqueness. These appended names shall be called the compound identifiers from hereon. Since only the mass spectrum is used for identification, several peaks may be identified as the same compound (best spectral match). However, this is not accurate, and these identities need to be confirmed. Confirming these identities is a labour-intensive process, which is simplified in this instance by only checking the statistically significant compounds and assigning metabolite names based on the visually inspected mass spectrum to these (explained in Section 5.4).

4.4.7 Normalisation of data

Due to instrument variation that might be introduced at various stages of the analysis (e.g. injection volume, column conditioning, inlet liner cleanliness, tune settings between batches), the data needs to be normalised to reduce this variation. Various methods exist for normalisation. When working with urine, samples are normalised to the creatinine concentration, in cellular and tissue metabolomics, mass, DNA or protein concentration is used for normalisation. Water homeostasis in humans is tightly regulated by adjusting its reabsorption in the kidneys. Therefore, the amount of water in the blood remains relatively constant (280 to 288 mOsm/kg) (Zerbe *et al.*, 1991). For this reason, the concentration of analytes in serum/plasma does not differ much between individuals and no sample-specific normalisation is implemented.

MetaboAnalyst (Chong *et al.*, 2018) provides several normalisation options including normalisation by sum, median, reference sample, pooled sample, reference feature and quantile normalisation. The method used for the extraction of these samples (explained in 4.3) includes internal standards which can be used for normalisation. Three normalisation methods were evaluated using CV curves, namely normalisation by 3-PHB only, normalisation by the sum of the areas of all four internal standards and normalisation using mass spectrum total useful signal (MSTUS). Figure 4-3 shows the CV-curves for the three normalisation strategies compared to the peak areas. From this figure, it is clear that all normalisation methods decrease the variability in the data similarly. For this reason, it was decided to normalise with the 3-PHB internal standard as per the default method.

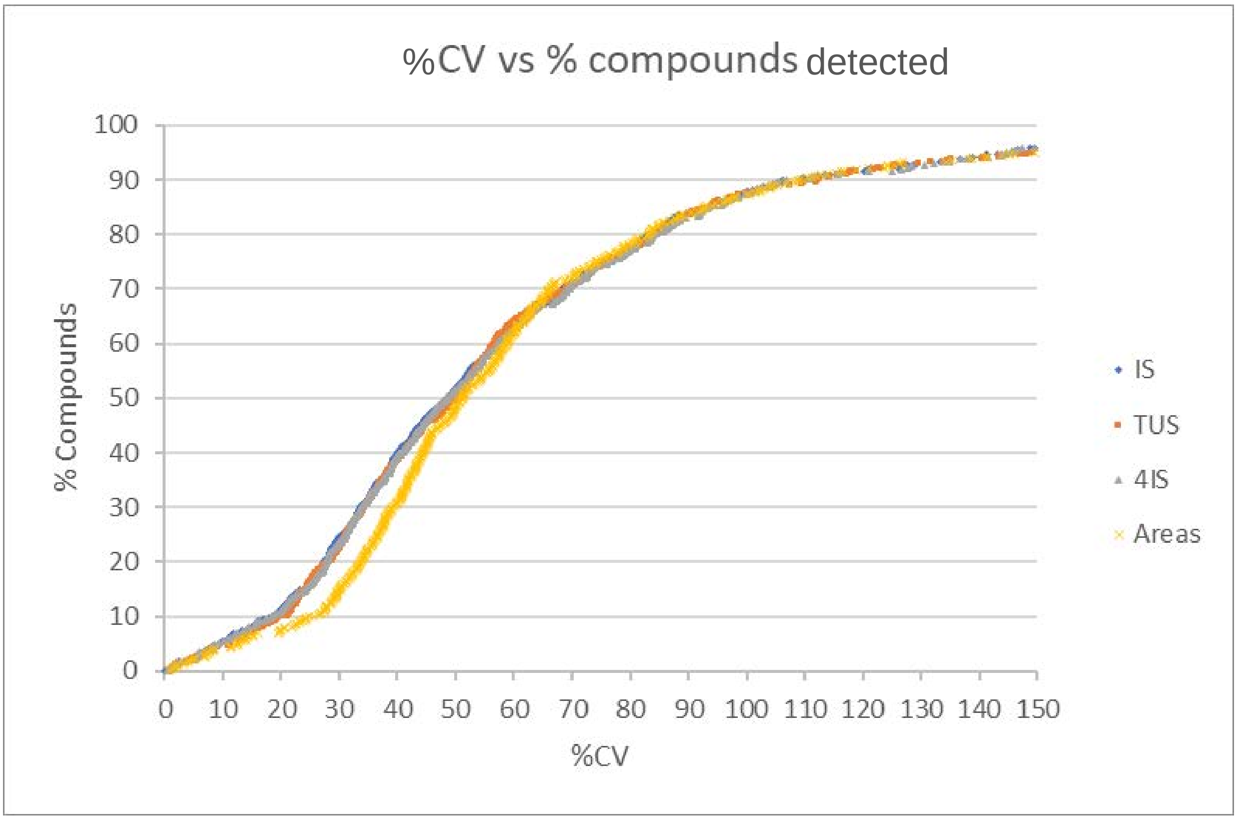

Figure 4-3: CV-curve of different normalisation techniques compared to unnormalised data (represented by the peak areas).

Since the concentration of 3-PHB is known, relative concentrations of other compounds could be estimated with the following formula:

$$C_A = \frac{A_A}{A_{IS}} \times C_{IS}$$

C_A represents the estimated relative concentration, A_A the area of the analyte to be quantified, A_{IS} the area of the internal standard peak and C_{IS} as the concentration of the internal standard. To calculate the concentration of 3-PHB after 50 µl plasma is mixed with 50 µl IS, the moles of 3-PHB in 50 µl of internal standard is divided by the total volume of 100 µl. Therefore, the concentration of 3-PHB is effectively halved. The addition of IS also dilutes the sample, but because equal quantities of internal standard and sample are added, dilution of the sample and internal standard cancel each other out and 3-PHB concentration can be taken as the stock concentration which is 36µg.ml^{-1}. This concentration is converted to moles per millilitre with the following simplified formula:

$$C_{mole} = \frac{C_{mg} \times 10^{-3}}{M}$$

With C_{mole} as the concentration expressed in moles per millilitre, C_{mg} the concentration expressed in milligrams per millilitre and M the molar mass of the compound of interest (3-PHB 164.2 grams per mole in this case). The resulting concentration of a compound with the same peak area as 3-PHB is then 219.2 mmol.l^{-1}.

4.5 Overview of the applied statistical approach

In short, the peak areas were quantified relative to the internal standard. After that, the samples were split into three different datasets to comply with the respective sub-studies. Pre-processing and quality assurance was performed on each dataset. The data was quality assured and batch corrected. Due to the size of the cohort, groups for some sub-studies were still quite small. For consistency, it was decided not to remove any outliers for any of the datasets representing the respective sub-studies. Parametric (independent t-test, Cohen’s d, PCA and PLS-DA) statistics was applied to the transformed data, and non-parametric statistics (Mann-Whitney test and its effect size) was applied to the untransformed data. Figure 4-4 shows the sequence of the statistical approach.

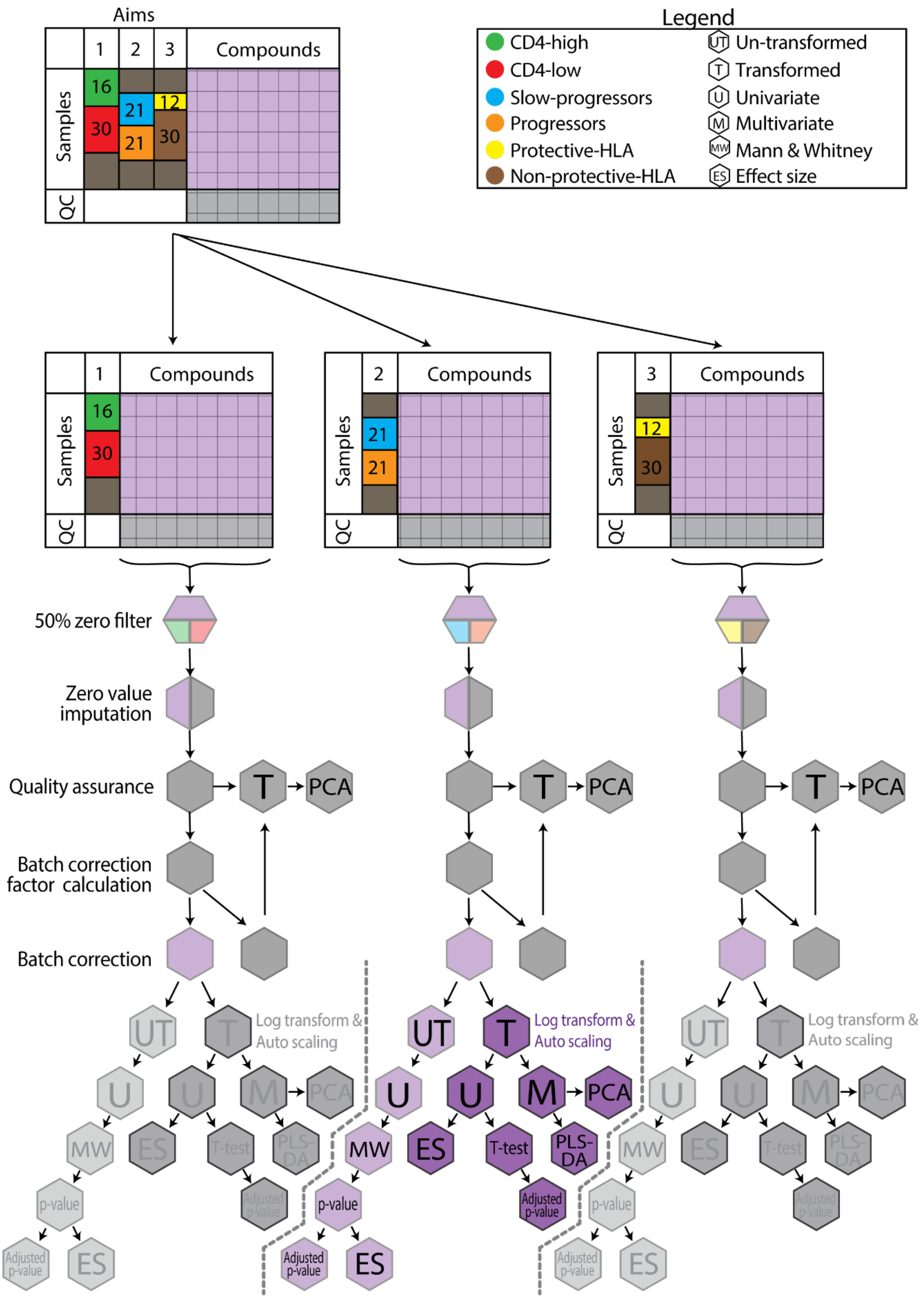

Figure 4-4: **Graphical representation of the series of statistical methods applied to our dataset (statistical methods are repeated for each dataset, for simplicity, only sub-study two is shown in colour).**

4.5.1 Pre-processing

Detection and identification of the peaks followed the analysis of the samples. Following this, the peak areas were aligned into a numerical area matrix of samples by compounds by aligning the peaks across all samples as described in 4.4.6. Pre-processing involves formatting the data from the raw output and performing normalisation and data clean-up steps to eventually obtain an interpretable dataset. Raw exported data contains various parameters such as retention times and quantification mass, in addition to the peak area. The peak area is used for normalisation and clean up, while other columns aids in confirming the identities of statistically significant compounds. The best spectral match from the library is assigned as the compound name. Multiple peaks were falsely identified as the same compound. Therefore, the compound name (name of the best match in the spectral library), unique mass (a unique mass fragment) and the two retention times were appended as the compound identifier. From this identifier, a compound can easily be found and validated on the chromatogram.

After assigning the peak identifiers, the internal standard was used to quantify the areas according to 4.4.7. The quantified data specific to each sub-study was then divided into separate datasets. Each dataset was individually subjected to the 50% zero filters. The 50% zero filter highlights those compounds detected in less than 50% of the samples of a specific group indicating that the compound might not be of much value as its concentration is too close to the detection limit of the instrument or the variation within a group is too much causing fluctuations over and below the detection limit. The compounds detected in less than 50% of samples of both groups were excluded from the dataset. After that, missing values were imputed by replacement with small random values (<½ of the minimum value in the dataset) mimicking the tails of a distribution. Zero value replacement makes the dataset compatible with the mathematical equations used in the statistical methods. Several equations require the same number of observations for each compound and each sample to work (no missing values), while others return errors when presented with data containing zero values. Therefore, neither missing or zero values can occur in the dataset.

4.5.2 Quality assurance

The analyst and machine repeatability, as well as the per sub-study quality assessment, was performed of the data.

4.5.2.1 Repeatability

Repeatability of both the analytical method and the machine was not performed before as this aspect has been previously reported by individuals from our laboratories (Olivier & Loots, 2012).

To ensure the trustworthiness of the data, machine (CV in quality control samples [QC-CV] within a batch) and analyst repeatability (QC-CV over batches) were assessed on the QC samples run throughout the analysis. This is represented as QC-CV plots. A QC-CV plot ranks all detected compounds in ascending order based on their relative standard deviation (RSD) or the coefficient of variance (CV). The rank number is then plotted against the CV. A lot of information can be obtained from such a plot, especially when comparing QC-CV plots of batches. Firstly, the shape of the plot gives an overview of the analytical error in quantification of compounds. A hyperbolic shaped curve with a sharp increase in rank number at a low CV indicates that the analytical error was minimal. A sigmoidal shaped curve indicates an analytical error that is relatively constant over a range of compounds but low or very high for other compounds. This type of curve is normally indicative of an outlier in the data. A single QC sample was prepared per batch but injected multiple times throughout the batch. Variation in the areas of the compounds between these injections is used to determine machine repeatability. The analyst repeatability is dependent on machine repeatability as variations in machine repeatability will inherently affect the outcome of analyst repeatability.

4.5.2.2 Transformation and per sub-study quality assurance

Data transformation involves adjustment of the physical data to achieve normality by applying an equation to every value in the dataset. Log transformation and auto-scaling were used to transform the data to normality. Log transformation is used to transform the positive skew occurring in metabolomics data towards normality by applying the $\log_{10}$ equation to all concentrations in each dataset (Field, 2000). Subsequently, the data was auto scaled, according to van den Berg *et al.* (2006). In short, auto-scaling centres the data around the dataset mean and divides by the standard deviation of each compound as the scaling factor (van den Berg *et al.*, 2006).

To assess the quality of the data in each sub-study, the data was transformed and auto scaled (explained in 4.5.5) where after, a PCA plot was used to assess the distribution between QCs and samples. A PCA is a simplified unsupervised representation of multidimensional data into components which maximises separation due to variation between samples (Barnes *et al.*, 2016; Xia & Wishart, 2016). A PCA, therefore, shows the maximum variation between samples in the first components irrespective of group labels. After the PCA model is drawn, the class labels are applied, and confidence intervals are drawn. Since group labels are not considered in the model and only applied afterwards, overlapping confidence regions in a PCA indicate that the group labels do not contribute to variation whereas separation shows that the group labels do in fact contribute to the variation between samples. The QC samples which should be representative of the samples analysed are expected to cluster together closely in the middle of the samples with a smaller confidence interval than the samples. A batch effect would portray as a separation of the QCs and samples of one or more batches from the rest. If evident, the batch effect was corrected as explained in 4.5.3. Upon sufficient PCA overlap of QC samples, as well as tight grouping centred within the samples, the untransformed and transformed datasets were subjected to further uni- and multivariate statistical procedures.

4.5.3 Batch correction

The batch correction was applied as described by (Van Der Kloet *et al.*, 2009). Briefly, the QC samples were used to determine which batch had the lowest average intra batch variation (best machine repeatability). This batch was not corrected. However, all other batches were corrected by adjusting their means to this specific batch. The correction factor was determined for each compound in each batch according to the following formula:

$$cf_{b,c} = \frac{\overline{C_{b,c}}}{\overline{C_{x,c}}}$$

With $cf_{b,c}$ the correction factor for compound c in batch b, $C_{b,c}$ the mean concentration of compound c in batch b and $C_{x,c}$ the mean concentration of compound c in the batch x, the batch being corrected to.

For all samples including the QCs in the respective batch, the concentration was divided by the correction factor, thereby adjusting the means of the QCs to be the same as that of the batch with the lowest average intra sample variation according to the following formula:

$$C_{r,s,b,c} = \frac{C_{s,b,c}}{cf_{b,c}}$$

With $C_{r,s,b,c}$ the corrected concentration for sample s compound c in batch b, $C_{s,b,c}$ the uncorrected concentration of sample s compound c in batch b and $cf_{b,c}$ the correction factor for compound c in batch b.

The batch correction was evaluated by assessing PCA plots of the QCs versus the samples (post-transformation). Effective correction should minimise the 95% confidence interval of the QCs as well as pull these closer together (closer to the batch corrected to).

4.5.4 Statistical methods on non-parametric (untransformed) data

Univariate statistics sequentially performs the specified statistical test for each compound across its observations in the compared groups. The univariate methods used included the Mann-Whitney test and its associated effect size.

The Mann-Whitney test is the non-parametric equivalent of the parametric independent t-test (described in the following heading) which is applied on transformed data. The principle of this test lies in the per group summation of ranks of the concentrations for each compound ranked from low to high and the difference rank-sum between the groups. The lowest rank-sum is called the test-statistic. The mean and standard error can then be calculated from the test statistic using the sample sizes of each group. Together, they are used to calculate a z-score which quantifies the group's distance from the population mean. An absolute z-score of greater than 1.96 is significant (Field, 2000). The distribution of z-scores is used to calculate the p-value, which is subsequently adjusted for multiple testing. This is done to reduce the probability of a false discovery (concluding that a compound differed significantly when it was just random chance). The Benjamini & Hochberg adjustment is applied to control the false discovery rate by adjusting the p-values (Benjamini & Hochberg, 1995).

The effect size (ES) explains how many standard deviations the means of the two groups are from each other. Dividing the test statistic with the total rank-sum gives the ES (Field, 2000). Effect sizes of 0.2,0.5,0.8 and 1.3 indicate small, medium, large and very large effect sizes, respectively (Sullivan & Feinn, 2012).

4.5.5 Statistical methods on parametric (transformed) data

The transformed and scaled concentrations (parametric data) of each compound (univariate methods) was compared between the respective groups by applying the effect size and independent t-test. Cohen's d quantifies the effect of the difference between the means (Cohen, 1988). The independent t-test compares the means of the two groups and determines if they are significantly different. This is done by testing the null hypothesis (means are the same) against the alternative hypothesis (differing means) and determining a significance level for which the alternative hypothesis will be accepted. This significance level is generally set at <0.05 for statistically significant differences in the means. Before comparing the test statistic with the significance level, the p-values are adjusted as per the Benjamini & Hochberg method defined in 4.5.4.

PCA analysis (explained in 4.5.2.2) was performed to assess the natural variation in the samples and whether the group labels of the respective sub-studies were attributed to this variation. Separation of the confidence regions would indicate that the group labels (e.g. progressor vs non-progressor) are the main factors influencing variation in the samples. Another multivariate method, PLS-DA analysis, was also performed on each dataset. PLS-DA is similar to a PCA in that it simplifies multidimensional data by variance in the samples. However, PLS-DA maximises the covariance between the compounds and the class labels, thereby maximising the separation of samples based on group labels (Barnes *et al.*, 2016; Xia & Wishart, 2016). Since PLS-DA is a multivariate statistical model, an accurate description of classification is dependent on the ratio between samples and compounds. For an accurate description of classification, the number of samples needed is exponentially more than variables (compounds) measured (Westerhuis *et al.*, 2008). In metabolomics, the ratio is often opposite with much more compounds than samples. This can lead to models that do not describe the classification accurately but rather by chance. To assess the accuracy of the model, it needs to be validated. Although various validation methods exist, leave-one-out cross-validation (LOOCV) was applied to all PLS-DA models. The cross-validation Q^2 value indicates the prediction accuracy of the PLS-DA by separating the data into a training and validation set. In this method, the validation set is a single sample while the training set includes all other samples. A PLS-DA plot is drawn with n-1 training samples whereafter the validation sample is placed in the model, and the classification (group to which it belongs) is predicted. The prediction accuracy is calculated from the number of misclassifications out of all the classifications (Westerhuis *et al.*, 2008). LOOCV returns two values, R^2 and Q^2. The R^2-value represents the prediction accuracy (distance between classes) of the complete model, while the Q^2-value represents the assessment of the models' predictive relevance (F. Hair Jr *et al.*, 2014). R^2 values greater than 0.67, 0.33 and 0,19 are suggested to indicate substantial, moderate and weak predictive accuracy (Peng & Lai, 2012). A Q^2-value close to the R^2-value indicates that the model works independently and would be reproducible.

4.5.6 Validating compound IDs

All peaks were identified by their identifier, as explained in 4.4.6. The spectral match for identification leads to a level 3 identification, as explained in 2.5.2. The spectral match assigned by computer algorithms is not always accurate. Fortunately, the top 10 spectral matches are returned for analyst comparison of the peak spectra. The labour-intensive process of spectral comparison is not performed for all compounds, but only for those found to be statistically significant.

An increase in the identification confidence of these compounds can be achieved through metabolite-metabolite correlation analysis (MMCA) (Madhu *et al.*, 2015). The concentrations of the significant metabolites in all samples were correlated with each other using Spearman correlation. Compounds that have significant positive correlations with each other are likely related. A biological link in addition to correlation increases the identification confidence even more.

RESULTS AND DISCUSSION

5.1 Quality assurance

Before any of the analytical data is used for biomarker identification, it is crucial to assess the quality of the analysis and the data generated. One way of achieving this is by evaluating analytical data acquired from the QC samples using a QC-CV curve. The variation of the data when comparing multiple injections of the same extracted QC sample represents machine repeatability which is shown per batch in Figure 5-1 (%CV calculated from peak areas).

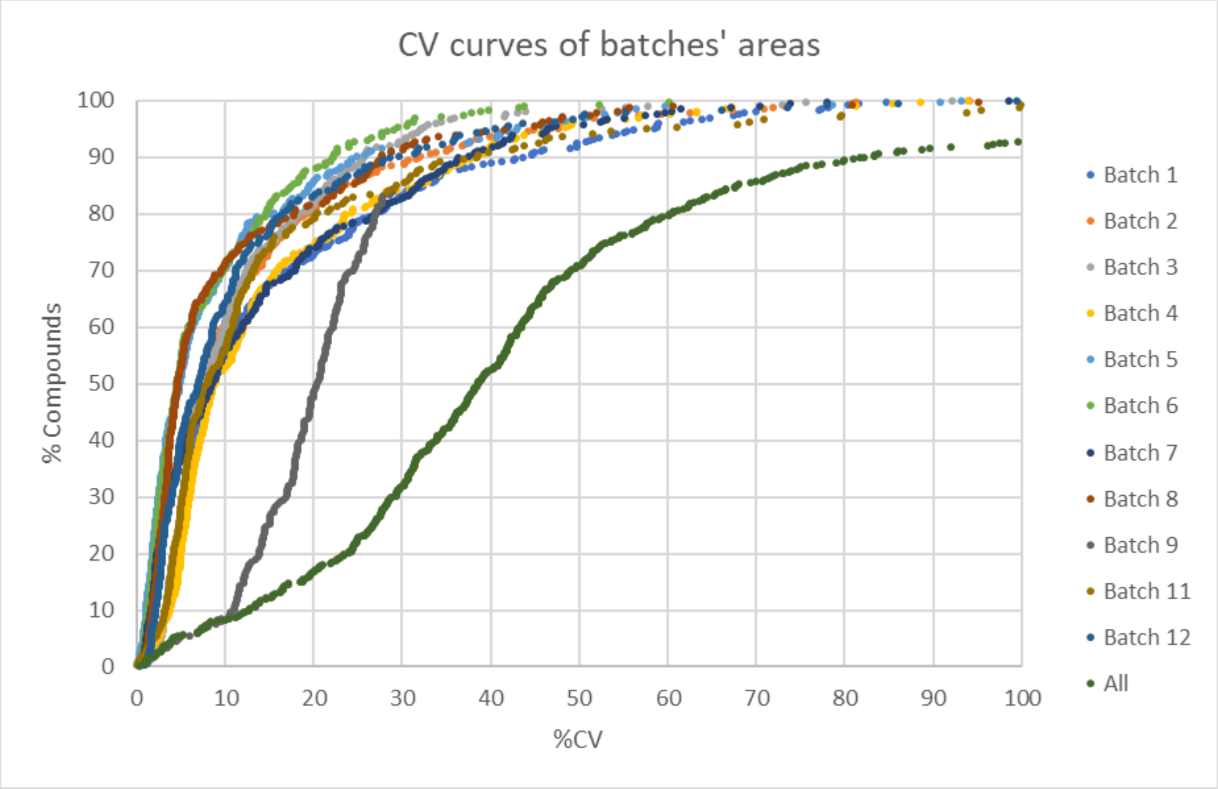

Figure 5-1: QC-CV curves of individual batches and all of the batches run together.

A total of 210 peaks were detected and were issued peak identifiers, as described in Section 4.5. The CVs of the analysed compounds over individual batches are plotted against the percentage of detected compounds in that batch. The results can be seen in Figure 5-1. The analyst repeatability, in turn, is assessed by comparing the compound CVs of the different QC samples extracted and analysed between the batches indicated as “all” in Figure 5-1. From Figure 5.1 it can be seen, that 90-98% of the detected compounds had CV’s of below 50% when assessing the machine repeatability, and 72% of analytes detected in QC samples had a CV of under 50%, when assessing analyst repeatability.

Considering the QC-CV curves of the individual batches, batch 9 appears as an outlier as seen by a right shift of the curve. Batch 9, however, was smaller than the other batches with only 3 QC sample injections. The third QC clogged the syringe and resulted in extremely low intensities across all peaks. This QC sample was excluded from the data set. The remaining two QC samples were used to re-calculate the CV. Since CV calculation includes the formula for standard deviation in which the number of samples is an important determinant, having only two QC samples adversely affected the CV values, but not the overall repeatability. PCA plots later confirmed that the QC samples of batch 9 overlapped with that of the other batches (Figure 5-2), indicating that no batch effects are present and the data is of good quality.

The QC-CV curve designated “all”, representing analyst repeatability has an even greater shift than batch 9 to the right. This curve is however not a pure representation of analyst repeatability, but a compounded representation of analyst, machine and extraction repeatability. The factors influencing this repeatability include retuning of the MS between batches, slight variations in temperatures during drying and derivatisation, residual moisture after drying and nitrogen flow during the drying of samples. These factors, in addition to analyst variability, can lead to rather large variations between batches. Hence the intra batch machine repeatability was greater than the between batch repeatability, as is generally expected.

Figure 5-2 shows the PCA plots of QCs versus sample data before and after batch correction, for sub-studies 1 to 3. At a glance, all the before and after batch-corrected PCA plots looked the same. This is expected since most of the compounds were retained in all datasets, and only a few were excluded by applying a 50% zero filter. Closer inspection reveals slight changes in the positions of some samples when comparing the datasets. Before the batch correction, a batch effect is evident, with the QCs and sample data from batch 11 and 12 presenting a left shift. This batch effect can be seen throughout all three datasets. After correction, the QCs and sample data cluster together with the QC points centred in the middle of the samples. Unfortunately, batch correction had an adverse effect on batch six which can be seen as a right shift of batch 6’s samples. Normally this batch would have been excluded from further analyses. However, we decided to keep this batch due to the small sample size of our cohorts in the respective sub-studies. All the statistical analyses described, under Section 4.5 was however applied on an additional dataset where this batch was removed to confirm that the batch effect observed as a result of this batch, does not contribute to the statistical significance of markers.

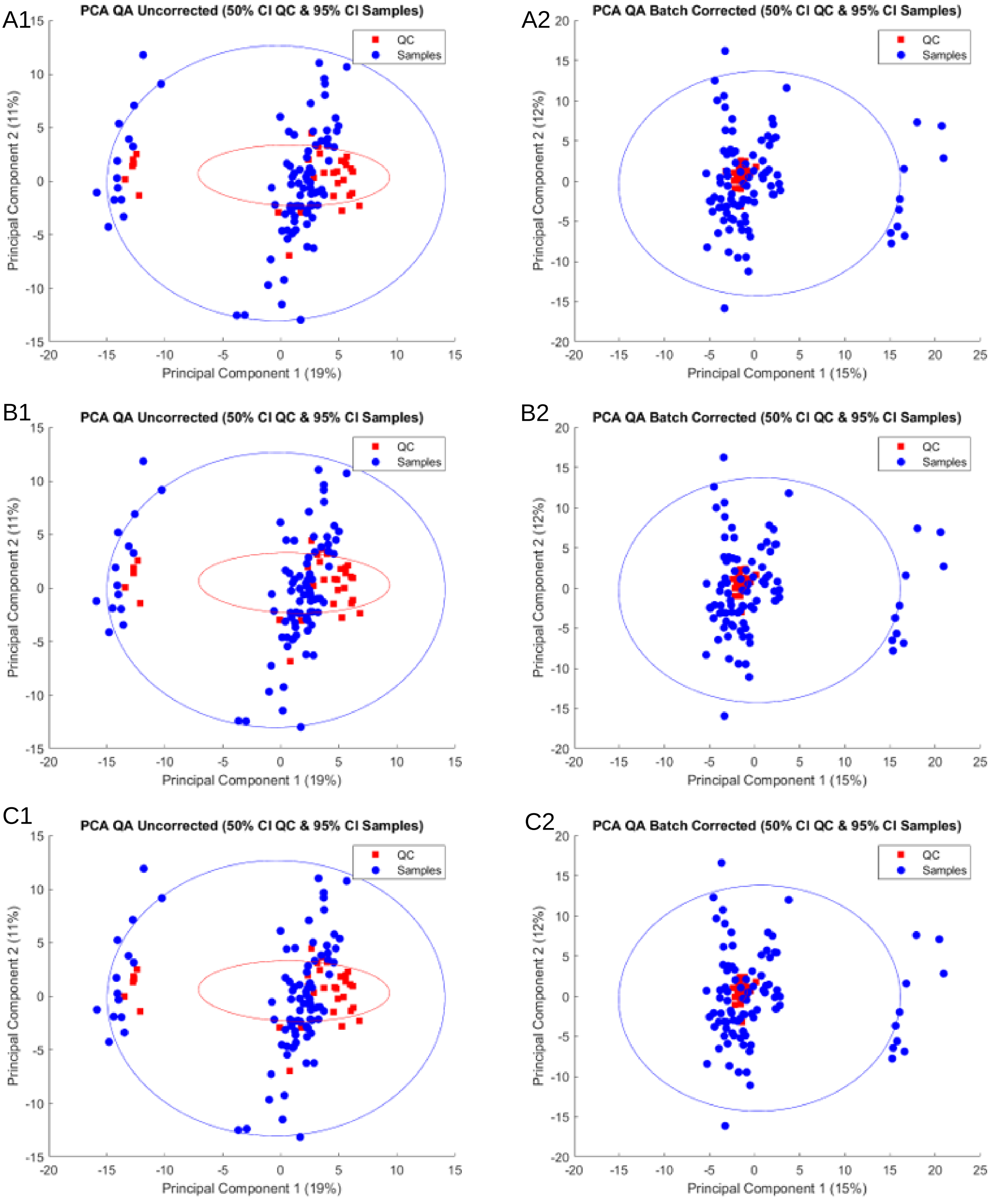

Figure 5-2: PCA plots of QCs versus samples after the 50% zero filters were applied to the data of sub-study 1 (A), sub-study 2 (B) and sub-study 3 (C). Datasets before (1) and after (2) batch correction are shown.

5.2 Multivariate statistical results

When considering the demographics and clinical data of the participants used in the respective sub-studies, stratification of samples shows the classification parameter to significantly differ between the groups. It is thus expected that one will see a statistically significant difference in the metabolite profiles of the respective groups. In each stratification, the classification parameter used was previously shown to be linked to HIV disease progression.

Statistical methods were applied as described in Section 4.5 in the sequence summarised in Figure 4-4. In short, the peak data was quantified relative to the internal standard, whereafter the samples were split into three different datasets to comply with the respective sub-studies. The data was pre-processed, and quality assured. The data was batch corrected whereafter the untransformed data was transformed, and multivariate (PCA and PLS-DA) statistics applied.

5.2.1 PCA results

Figure 5-3 shows PCA plots of groups of samples pertaining to each sub-study with and without the outlier batch, batch 6. The ellipses for each group indicates the 90% confidence region (area in which 90% of samples can be expected to occur). Considering batch correction, the exclusion of batch six does not change the amount of overlap in confidence regions between samples grouped according to the respective sub-studies. The percentage variance explained by the principal component (PC) 1 and PC2 in each case is similar before and after batch correction. Although between-group differences and within-group differences dominate the PCA, they are still considered small, implying homogeneity in the metabolic profiles of the samples. The samples were all collected from participants in the Umlazi area implying that the individuals are from an area of similar income status and culture. We, therefore, anticipate that their lifestyle and primary diet will not differ vastly, thus having little impact on the measured intra-group metabolic profiles.

Figure 5-3-A shows the PCA plot of the CD4-high versus the CD4-low groups. The massive overlap between these 90% confidence interval ellipses indicates that the highest variation in the data is not attributed to the CD4 count. The separation in principal component 1 in Figure 5-3-A1, however, shows that the highest variation is due to the right shift of batch six after correction. After batch six was removed (Figure 5-3-A2), the same trend still exists, with a downward shift of the CD4-low group observed. This indicates that CD4 count attributes only slightly to the variance in metabolite profiles.

Figure 5-3-B shows the PCA plot of samples of progressors versus non-progressors based on the correlation between CD4 count and time, as explained in Section 3.7. Though there is an overlap in the clustering of progressors and non-progressors, the samples within the respective groups show to be homogeneous, as seen in Figure 5-3-B1. After the removal of batch six, the confidence intervals overlap completely, indicating that CD4 count correlation with time does not explain the variance in metabolite profiles.

Figure 5-3-C shows the PCA plot of samples from participants with protective versus non-protective *HLA-B* alleles. Here the 90% confidence intervals once again overlap; however, the removal of batch six slightly decreases the overlap indicating that *HLA-B* alleles may explain some variance in the metabolite profiles of these samples.

The lack of separation of the samples in the respective sub-studies in Figure 5-3 is unexpected since CD4 T-cell depletion, the rapid rate of loss of CD4 cells and the presence of non-protective *HLA-B* genotypes have previously been linked to increased HIV-induced immune activation (Mellors *et al.*, 2007; Sousa *et al.*, 2002; Uribe *et al.*, 2004). This is, in turn, associated with higher inflammatory cytokine production and secretion and subsequent metabolic changes (Aounallah *et al.*, 2016; Fitzpatrick & Young, 2013; Nixon & Landay, 2010; Sitole *et al.*, 2013). Venner *et al.* (2016) investigated differences in disease progression based on HIV-subtype. The HIV-1 subtype C cases progressed slower than those infected with HIV-1 subtype D. This was explained by the characteristically lower replication capacity of subtype C when compared to that of subtype D. This comparatively lower subtype C replication capacity, and assumed lower metabolic turnover, may reduce the metabolite variation within the samples of a particular country/population. Indeed, the samples studied here were of subtype C origin, partially explaining the lack of metabolic variation within our cohort. Despite its lower replication rate, however, HIV-1 subtype C is known to be the virulent dominant strain globally, and hence the authors conclude that the expansion and dominance of subtype C globally, may be due to the longer periods of asymptomatic infection associated with this subtype.

Metabolic differences can, however, not be disregarded only based on a PCA not separating. Since PCAs are multivariate projections of data, these models become exponentially more complex, as more variables (compounds) are included. Furthermore, PCA plots do not consider class labels and aim to find the highest variation between the samples compared. This can be seen in the batch effects that are detected in batches 11 and 12 in the uncorrected data and then batch six after correction. We expected the non-relevant biological variance to be small, but it appears that it might play a larger role here than expected therefore diet, genetic factors, lifestyle factors, etc. might mask the minor changes in metabolites due to differences in HIV disease progression or CD4 counts. This does not mean that there are no metabolite differences; it only means that the metabolite differences due to differential HIV disease progression or CD4 counts cause less metabolic variation than other variation in the sample data.

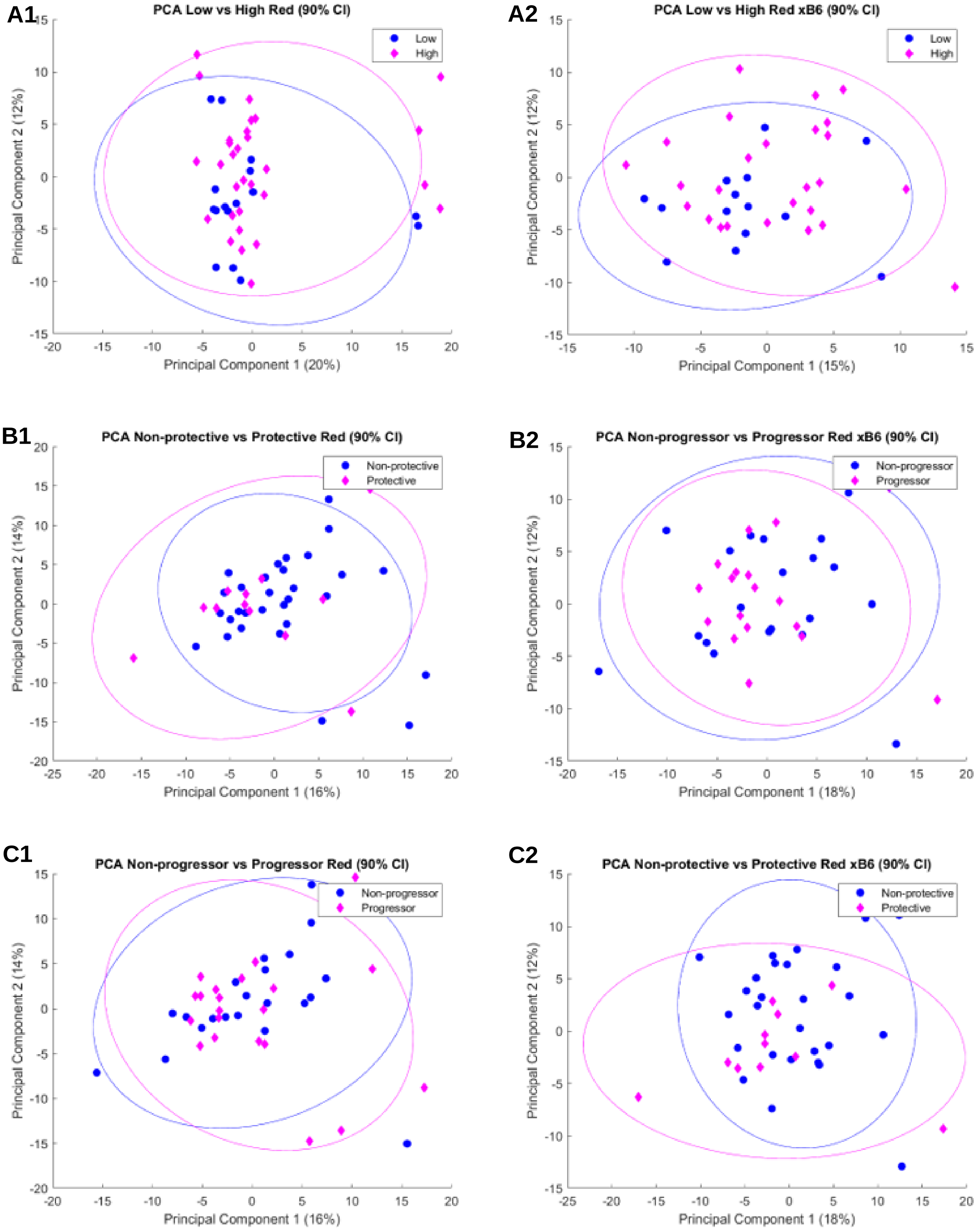

Figure 5-3: **PCA plots of the respective groups compared in sub-study 1 (A), sub-study 2 (B) and sub-study 3 (C) including batch 6 (1) and excluding batch 6 (2).**

5.2.2 PLS-DA results

Figure 5-4 shows the PLS-DA plots of the groups of each sub-study with and without the outlier batch, batch six. Groups from the respective sub-studies separated as expected in a discriminant analysis. Patient samples stratified according to CD4-high and low counts separated along the first component describing 52% of the variation. Patient samples stratified as progressors or non-progressors also separated along component one, and explained 45% of the variance. Participants with protective and non-protective *HLA-B* alleles similarly separated in the first component and explained 41% of the variance by the group labels. Greater separation was seen between cases stratified according to CD4 count than a change in those stratified by CD4 change in time, while grouping according to *HLA-B* alleles reflect the least separation in component one. This might be attributed to greater homogeneity in the mid-CD4 samples stratified according to change in CD4 over time and *HLA-B* alleles, respectively. Discriminant analysis is prone to overfitting. Thus the model needs to be cross-validated. Table 5-1 shows the cross-validation R^2 and Q^2-values of the PLS-DA plots in Figure 5-4. The interpretation of these values was discussed in 4.5.5. Briefly, an R^2 value greater than 0.67 is indicative of substantial prediction accuracy, which must be validated by a Q^2 value close to the R^2 value. The R^2 value for sub-study 1 is substantial while moderate for sub-studies 2 and 3, implying relatively good prediction accuracy. The Q^2-values are all extremely weak, indicating that this model is very volatile and cannot be used to interpret HIV disease progression further than this cohort. Hence, this indicates sufficient separation in the PLS-DA models for the comparisons of the data for all three sub-studies, for this specific cohort, was achieved. That said, although the variable of importance on projection (VIP) value for each compound selected as a marker is indicated in the results, this was not used for biomarker selection. The VIP was only used to complement the other univariate statistics.

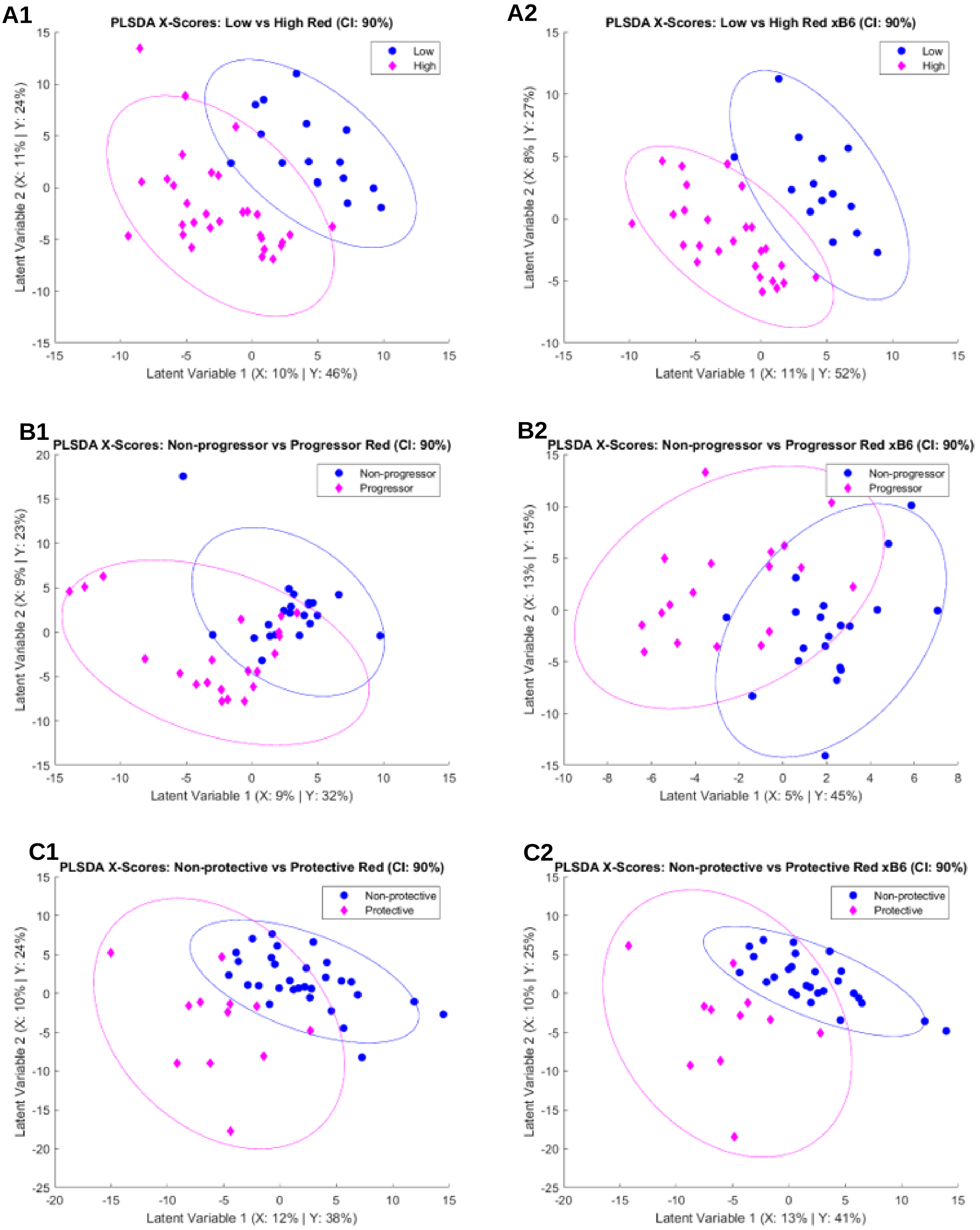

Figure 5-4: **PLS-DA plots of the respective groups compared in sub-study 1 (A), sub-study 2 (B) and sub-study 3 (C) including batch 6 (1) and excluding batch 6 (2).**

Table 5-1: R^2 and Q^2-values for PLS-DA plots

	With Batch 6		Without Batch 6	
	R^2	Q^2	R^2	Q^2
Sub-study1	0.70	0.08	0.78	0.06
Sub-study2	0.55	-0.95	0.61	-1.07
Sub-study3	0.62	-0.13	0.66	-0.12

5.3 Important compounds

Various statistical methods were applied to the data as described in Section 4.5, and summarised in Figure 4-4. In short, the internal standard was used to normalise data and quantify the compounds detected. The analysed data was split into three separate datasets containing information on only those samples needed to investigate the specific sub-study. Various pre-processing and quality assurance statistics were performed on the data. Batch correction was performed, followed by univariate analysis of the untransformed data using the Mann-Whitney test and its associated effect size. The batch corrected data was then transformed and univariate (effect sizes [Cohens' d] and independent t-test), and multivariate (PCA and PLS-DA) statistics applied. All p-values were adjusted for multiple testing, as described in Section 4.5.4.

Since the PLS-DA models failed to validate, univariate statistics were used to determine which compounds significantly differed when comparing the data collected from the sample groups associated with the respective sub-studies. The Venn diagram depicted in Figure 5-5 summarises the amounts and distribution of the significant markers across the sub-studies. Surprisingly, no metabolites were common across all sub-studies. This is likely due to the small number (n=4) of significant metabolites in sub-study 2. The biggest overlap in metabolites is between sub-studies 1 and 3, with 5 compounds common to both, 3 of which could be positively identified: isoleucine, methionine and glycine (Table 5-2).

After identifying the significantly altered compounds for each of the respective sub-studies, the compound identifier of the statistically significant compounds was translated into the metabolite name, HMDB name and HMDB ID, as shown in Table 5-2 and discussed in Section 5.4.

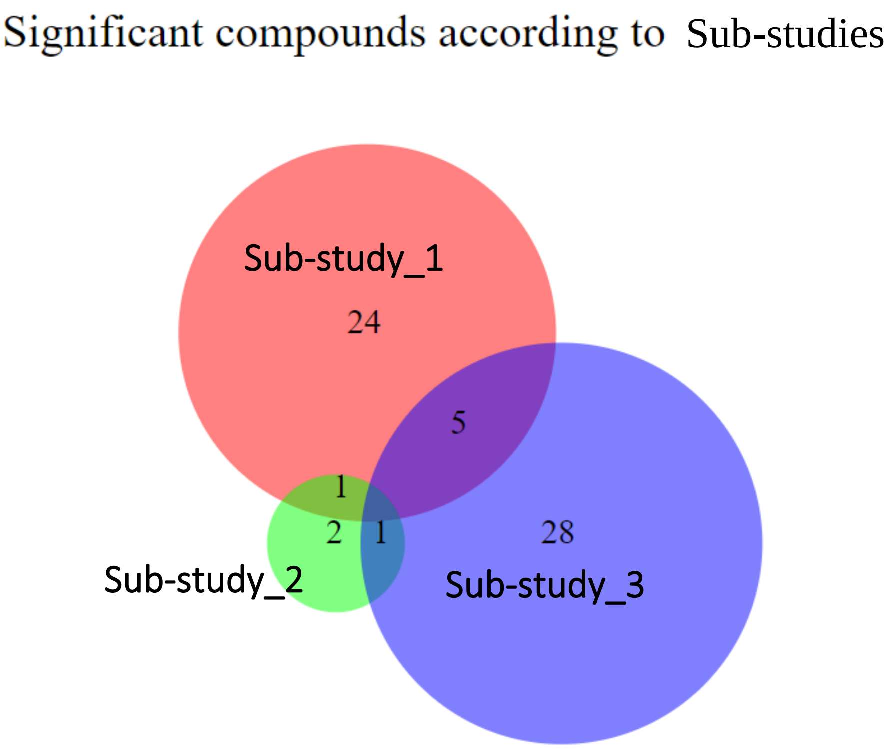

Figure 5-5: **Venn diagram showing the number of significant compounds identified in the respective sub-studies with a possible role in HIV disease progression.**

5.4 Confirming compound IDs

All GC-generated peaks have been identified up to this point by comparing their characteristic mass spectra to that of both commercially available and in-house compound mass spectral libraries, using their spectral library match, unique mass and retention times (termed the compound identifier as explained in Section 4.4.6). Figure 5-6, shows a confident match for L-tyrosine 3 TMS, as an example of only one of the many compounds identified in this manner. The confidence lies in the spectral difference between the peak and the library spectra. As seen in Figure 5-6B, the spectral difference is minimal, indicating an excellent match.

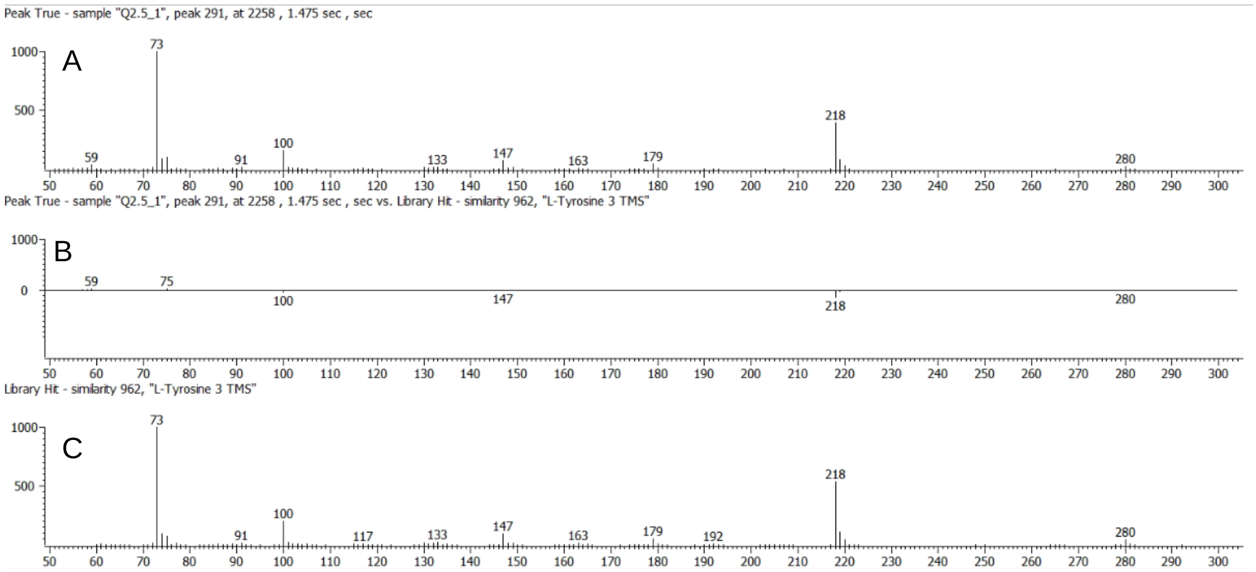

Figure 5-6: Example of peak confirmation of L-Tyrosine 3 TMS. (A) shows the peak spectra, (C) shows the library spectra and (B) the spectral difference.

Confidently matched compounds were translated to their underivatised metabolite name, as indicated in Figure 5-7 and subsequently to the HMDB compound name and identifier. Figure 5-7 shows how an identified compound is underivatised *in silico* to reveal the original metabolite before derivatisation. The derivatisation groups added to the parent molecule native to the original plasma sample are indicated in red (the trimethyl silane [TMS] group) and blue (the methoxime group). The figure below also shows that the compound ID changed from the International Union of Pure and Applied Chemistry (IUPAC) name *pentanoic acid, 2-(methoxyimino)-3-methyl-, trimethylsilyl ester* to the more commonly used biochemical name: isoleucine. This translation of compound IDs was performed for all derivatised compounds identified as significant when comparing the metabolomics data between the sample groups specific for each sub-study. The compound identifier and its allocated metabolite names are presented in Table 5-2.

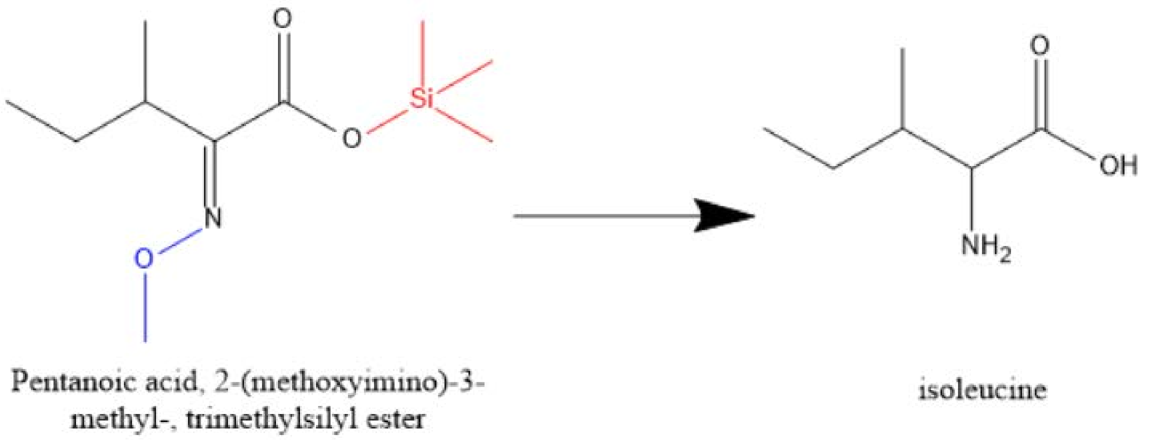
Pentanoic acid, 2-(methoxyimino)-3-methyl-, trimethylsilyl ester
isoleucine

Figure 5-7: **Example of the translation of a compound to a metabolite. Here the conversion of Pentanoic acid, 2-(methoxyimino)-3-methyl-, trimethylsilyl ester to isoleucine through changing of the derivatised groups is shown.**

For convenience, the rows of compounds in Table 5-2 were coloured according to the Venn diagram shown in Figure 5-5.

Table 5-2: **Table of compound names and their corresponding metabolite names, HMDB names, HMDB ID, and sub-studies in which they were significant. (Coloured according to Figure 5-5)**

Library hit Compound	Metabolite	HMDB NAME	HMDB ID	SUB-STUDY
4-Hydroxyproline 3 TMS_73_1594.58_1.38	4-hydroxyproline	4-Hydroxyproline	HMDB0000725	1
2-Ketoisocaproic acid mo-tms_73_985.98_1.82	2-ketoisocaproic acid	Ketoleucine	HMDB0000695	1
3,3-Dichloropropyne_73_1151.13_2.08	Benzene acetic acid	Phenylacetic acid	HMDB0000209	1
Analyte 279_73_1687.36_1.94	Unknown	Unknown	Unknown	1
Analyte 341_73_1994_1.41	Unknown	Unknown	Unknown	1
Benzene, 1,2,3-trimethyl-_105_478.12_1.01	Unknown	Unknown	Unknown	1
Dodecanoic acid, TMS derivative_117_1813.85_1.43	Dodecanoic acid	Dodecanoic acid	HMDB0000638	1
Glyceric acid, 3TMS derivative_73_1226_1.4	Glyceric acid	Glyceric acid	HMDB0000139	1
Glycerol 3 TMS_147_1115.3_1.27	Glycerol	Glycerol	HMDB0000131	1
L-Aspartic acid 3 TMS_73_1585.98_1.43	L-Aspartic acid	L-Aspartic acid	HMDB0000191	1
L-Isoleucine 2 TMS_158_1148_1.39	L-Isoleucine	L-Isoleucine	HMDB0000172	1
L-Leucine 2 TMS_158_1106_1.38	L-Leucine	L-Leucine	HMDB0000687	1
L-Phenylalanine 2 TMS_218_1760.08_1.68	L-Phenylalanine	L-Phenylalanine	HMDB0000159	1
L-Proline 2 TMS_142_1151_1.57	L-Proline	L-Proline	HMDB0000162	1
L-Tryptophan, 1-(trimethylsilyl)-, trimethylsilyl ester_202_2608.71_2.2	L-Tryptophan	L-Tryptophan	HMDB0000929	1
L-Tyrosine 3 TMS_218_2257.53_1.5	L-Tyrosine	L-Tyrosine	HMDB0000158	1
L-Valine 2 TMS_144_989_1.39	L-Valine	L-Valine	HMDB0000883	1
Methanol, TMS derivative_89_670.93_1.42	Methanol	Methanol	HMDB0001875	1
Nonanoic acid, TMS derivative_129_1285.36_1.48	Nonanoic acid	Pelargonic acid	HMDB0000847	1
Pentanoic acid, 2-(methoxyimino)-3-methyl-, trimethylsilyl ester_73_911_1.8	Isoleucine	L-Isoleucine	HMDB0000172	1

Library hit Compound	Metabolite	HMDB NAME	HMDB ID	SUB-STUDY
Pentanoic acid, 2-(methoxyimino)-3-methyl-, trimethylsilyl ester_73_955.87_1.84	Isoleucine	L-Isoleucine	HMDB0000172	1
Pyruvic acid, TMS derivative_117_1469_1.46	Decanoic acid	Capric acid	HMDB0000511	1
Ribitol, 5TMS derivative_73_1928.53_1.17	Ribitol	Ribitol	HMDB0000508	1
Uridine, 3TMS derivative_73_2917.6_1.78	Uridine	Uridine	HMDB0000296	1
1,2-Benzenedicarboxylic acid, butyl 2-ethylhexyl ester_149_2280.88_2.33	Unknown	Unknown	Unknown	2
D-Fructose, 1,3,4,5,6-pentakis-O-(trimethylsilyl)-, O-methyloxime_103_2164.77_1.23	Carbohydrate	Unknown	Unknown	2
1,3-Dioxolane_73_1213.85_1.1	Unknown	Unknown	Unknown	3
2-Aminomalonic acid, N,O,O,-TMS_147_1490_1.56	2-Aminomalonic acid	Aminomalonic acid	HMDB0001147	3
Analyte 148_73_926.58_1.1	Unknown	Unknown	Unknown	3
Analyte 173_116_1074.1_1.62	Serine	L-Serine	HMDB0000187	3
Analyte 187_155_1118.04_0.29	Unknown	Unknown	Unknown	3
Analyte 262_116_1601_1.77	Paracetamol	Acetaminophen	HMDB0001859	3
Analyte 280_166_1688.02_2.3	Unknown	Unknown	Unknown	3
Analyte 364_142_2077.62_1.28	Ornithine	Ornithine	HMDB0000214	3
Analyte 401_157_2182.96_1.33	Unknown	Unknown	Unknown	3
Analyte 540_73_3298.5_1.51	Glycerol monostearate	MG(0:0/18:0/0:0)	HMDB0011535	3
Benzoic Acid, TMS derivative_179_1049_1.97	Unknown	Unknown	Unknown	3
D-Glucose 5 TMS; 1 MEOX_147_2185.79_1.22	Carbohydrate	Unknown	Unknown	3
Erythro-Pentonic acid, 2-deoxy-3,4,5-tris-O-(trimethylsilyl)-, trimethylsilyl ester_73_1828.94_1.26	Arabinoic acid	Arabinonic acid	HMDB0000539	3
Glycine 3 TMS_174_1171.94_1.41	Glycine	Glycine	HMDB0000123	3
Glycine, di-TMS_102_794_1.45	Glycine	Glycine	HMDB0000123	3
Glycolic acid, 2TMS derivative_147_697.62_1.49	Glycolic acid	Glycolic acid	HMDB0000115	3
L-Isoleucine, TMS derivative_86_959_1.69	L-Isoleucine	L-Isoleucine	HMDB0000172	3
L-Lysine, 4TMS derivative_156_2236.65_1.27	L-Lysine	L-Lysine	HMDB0000182	3
N,O-Bis(trimethylsilyl)carbamate_100_667.97_1.58	Carbamate	Carbamic acid	HMDB0003551	3
Octanoic acid, 2-dimethylaminoethyl ester_58_2971.5_1.56	Octadecanoic acid	Stearic acid	HMDB0000827	3
Oleic Acid, (Z)-, TMS derivative_75_2383.51_1.46	Palmitelaidic acid	Palmitelaidic acid	HMDB0012328	3
Oxalic acid, 2TMS derivative_147_818_1.73	Oxalic acid	Oxalic acid	HMDB0002329	3
Pyruvic acid, TMS derivative_75_892.98_1.51	Heptanoic acid	Heptanoic acid	HMDB0000666	3
Ribitol, 5TMS derivative_73_1568_1.2	Unknown	Unknown	Unknown	3
Ritalinic acid, 2TMS derivative_156_1220_1.5	Pipecolic acid	Pipecolic acid	HMDB0000070	3
Serotonin 5 TMS_174_2935.69_1.61	Serotonin	Serotonin	HMDB0000259	3
Urea, 2TMS derivative_147_1054.15_1.85	Urea	Urea	HMDB0000294	3
Urea, 3TMS derivative_147_916.98_1.32	Urea	Urea	HMDB0000294	3
1,3-Dioxolane_73_2488.46_1.68	Unknown	Unknown	Unknown	1,2
1-Methyl-N,N-bis(trimethylsilyl)-4-[(trimethylsilyl)oxy]-1H-imidazol-2-amine_115_1633.6_1.67	Creatinine	Creatinine	HMDB0000562	1,3
2-Methyl-2-(trimethylsilyloxy)-1-(4-(2-(trimethylsilyloxy)ethoxy)phenyl)propan-1-one_131_1586_1.33	Unknown	Unknown	Unknown	1,3
Analyte 156_73_977.04_1.41	Glycine	Glycine	HMDB0000123	1,3
Ethanolamine 3 TMS_174_1093.86_1.29	Unknown	Unknown	Unknown	1,3
L-Methionine 2 TMS_176_1576.57_1.62	L-Methionine	L-Methionine	HMDB0000696	1,3
L-(-)-Sorbose, pentakis(trimethylsilyl) ether, methyloxime (syn)_103_2179.85_1.24	Carbohydrate	Unknown	Unknown	2,3

Since the same amino acid can be derivatised differently, and result in up to 3 different peaks at three different retention times, we can add their areas together. However, the identifications obtained for our study were level 3 identifications. When adding peak areas based on this identification level, mistaken identifications can lead to different compounds being summed and adversely affect further statistical results. For this reason, multiple derivatised peaks were analysed as individual compounds.

The reason for appending the unique mass and retention times to the identified compound name becomes apparent when considering, for example, the two compounds in red text in Table 5-2, that were matched to the same compound in the NIST library, but had different retention times. Figure 5-8 shows the spectral similarity between these two compounds. From the spectra, it is clear why they have been identified as the same compound. However, a 45 second retention time gap indicates that they are chemically different.

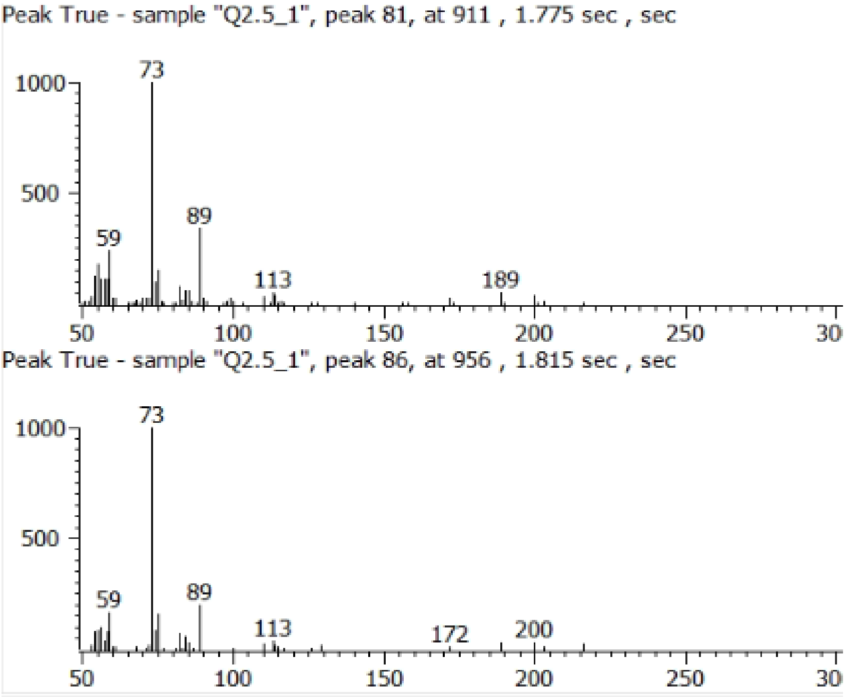

Figure 5-8: **Side-by-side comparison of the spectra of two statistically significant compounds (marked with red text in Table 5-2) eluting at different retention times.**

It is well known that derivatisation does not yield homogenous products, often resulting in a mixture of mono-, di- or tri-TMS derivatised compounds. The TMS groups chemically alter the compound, effectively changing its mass, polarity as well as functionality (Zarate *et al.*, 2017). Multiple peaks representing the different grades of derivatisation is therefore expected for some compounds, but their spectra are more often than not also slightly altered. Figure 5-9 shows the spectral difference between mono- and di-TMS L-leucine.

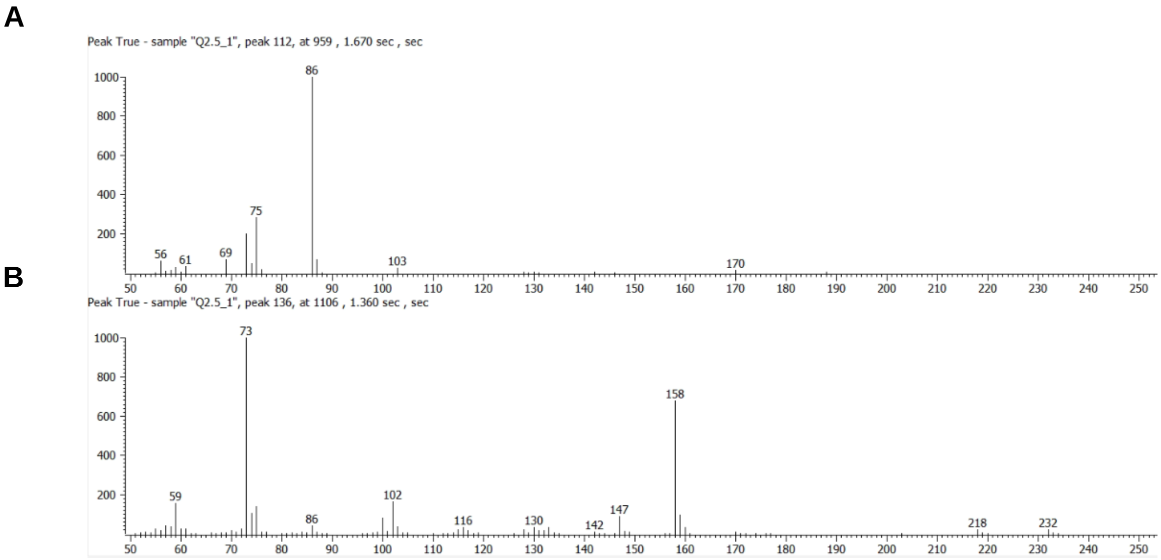

Figure 5-9: **Spectra from L-leucine (A) mono-TMS and (B) di-TMS**

Isomers are chirally different, and although the GC column was not chosen for chiral selectivity, there might be a retention time difference attributed to the compounds' chirality. Chirality can, however, not be resolved by spectra alone. Figure 5-10 shows the spectral similarity between the D- and L- isomeric forms of leucine.

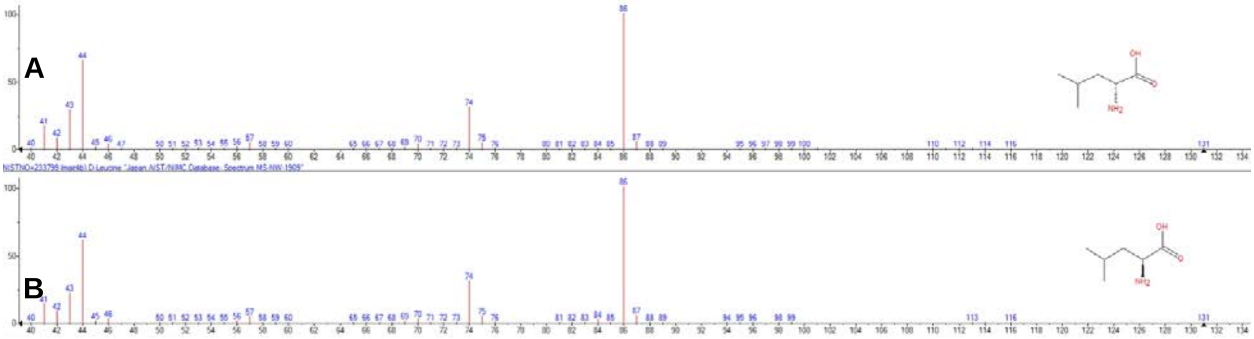

Figure 5-10: **NIST library spectra of (A) D- and (B) L-leucine**

Figure 5-11 shows the Spearman correlation coefficients of the significant metabolite-metabolite correlations analysis (MMCA) across all sub-studies. Metabolites 1 to 6 all had significant correlations with each other although metabolites 1 to 3 and 4 to 6 clustered together with correlation coefficients higher than 0.8. Since these metabolites were identified as branched-chain amino acids and their respective products according to spectra, the correlation between them increases the confidence that they are, in fact, all related to the branched-chain amino acid catabolism. Due to the spectral similarities, it is likely that the two compounds identified as isoleucine might be the keto acid catabolites of branched-chain aminotransferase (BCAT).

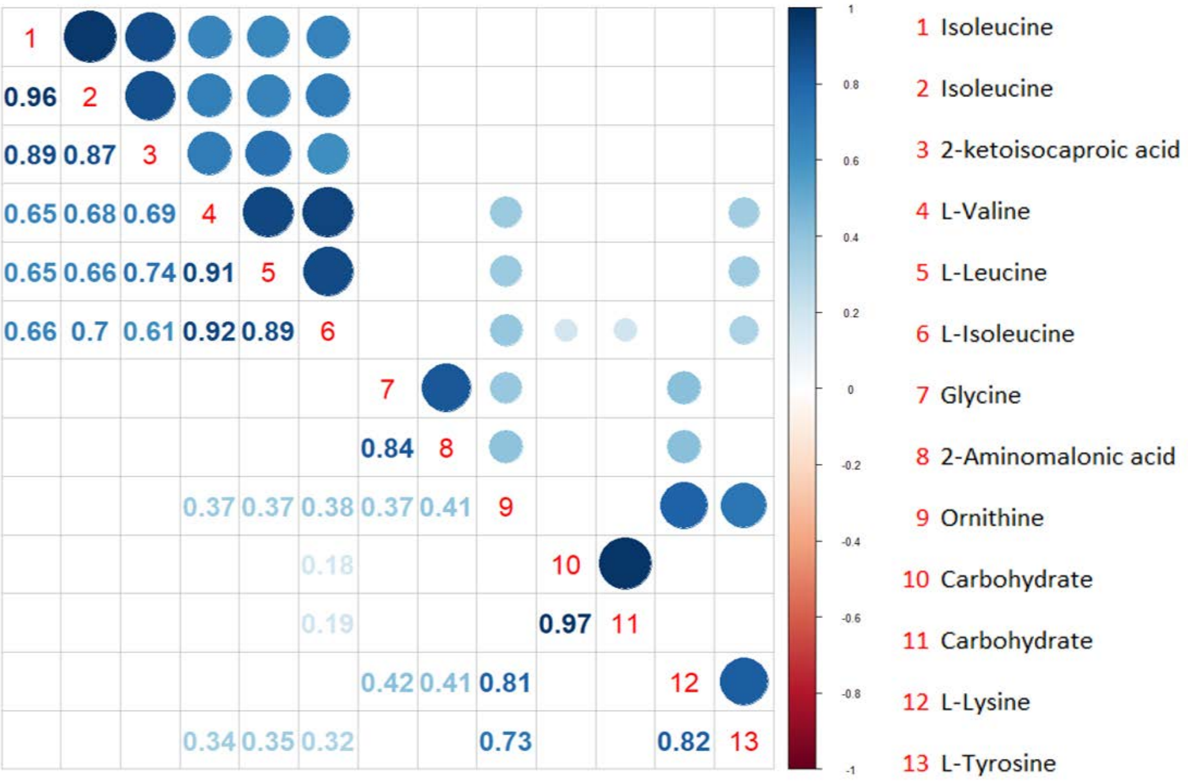

Figure 5-11: Metabolite-metabolite correlation analysis. Spearman correlation coefficient displayed numerically and graphically for significantly (p<0.05) correlating metabolites.

By now, the limitation of only being able to identify compounds based on mass spectra should be evident. However, since this is an exploratory, hypothesis-generating untargeted study, biological interpretations will be made on the assumption that the identification is correct, and provide suggestions as to which compounds could be analysed in a targeted metabolomics approach. Compounds named as “unknown” did not pass the visual spectral screening and remains to be identified at a later stage.

5.5 Metabolomics of HIV disease progression

The metabolites with significant differences when comparing the groups in the respective sub-studies along with their statistical results are indicated in Table 5-2. MetaboAnalyst (Chong *et al.*, 2018) was subsequently used for pathway analysis of the significant metabolites per sub-study and as a whole.

5.5.1 CD4-high versus CD4-low comparison

Table 5-3 lists the metabolites deemed significant between untreated HIV positive participants with high versus low CD4 counts. The colours of the means in the table show the increased value in red and the decreased value in green. Statistically significant (MW adjusted p-value ≤ 0.05; MW effect size ≥ 0.5; Unpaired t-test adjusted p-value ≤ 0.05; ES Cohen’s d-value ≥ 0.8) and slightly significant (MW adjusted p-value ≤ 0.1; MW effect size ≥ 0.3; Unpaired t-test adjusted p-value ≤ 0.1; ES Cohen’s d-value ≥ 0.5) metabolites between the CD4-high and CD4-low groups are coloured in lime and yellow, respectively. The PLS-DA variable importance in projection is also included, but not used for classification of the metabolites in this instance. The data is merely shown to illustrate that the multivariate analysis complements those metabolites found to be of significance through univariate tests. The most obvious observation is that participants with low CD4 counts had low concentrations of amino acids. This was paralleled by a general decrease in even chain fatty acids and carbohydrates. Nonanoic acid, an odd chain fatty acid was the only fatty acid which increased. Benzene acetic acid, L-aspartic acid and methanol were also increased in the HIV positive participants with low CD4 counts.

Table 5-3: Table of statistically significant metabolites measured between the samples of participants with high and low CD4 counts (AA = Amino acids, FA = Fatty acids, CA = carbohydrates, U = Unknown). Metabolites that increased are highlighted in red while those that decreased are highlighted in green.

		Untransformed						Transformed data		
		Mean(mmol/l)		Standard deviation		Mann-Whitney test		Effect size		PLS-DA VIP**
Metabolite	Class	CD4-Low	CD4-High	CD4-Low	CD4-High	Adjusted P-value*	Effect size	Adjusted p-value*	Cohen's d-value	
2-ketoisocaproic acid	AA	7.20	9.37	2.78	2.30	0.12	0.43	0.21	0.84	2.21
4-hydroxyproline	AA	4.08	5.28	2.63	1.95	0.22	0.38	0.38	0.64	1.78
Creatinine	AA	8.06	12.24	3.25	5.99	0.39	0.33	0.21	0.77	1.53
Glycine 3	AA	0.85	1.13	0.54	0.53	0.57	0.26	0.57	0.53	1.32
Isoleucine 1	AA	5.76	7.97	1.68	3.96	0.43	0.30	0.26	0.66	1.69
Isoleucine 2	AA	3.05	4.72	2.11	4.42	0.40	0.31	0.53	0.54	1.31
L-Isoleucine 1	AA	11.58	16.01	2.51	4.72	0.04	0.53	0.02	1.15	2.68
L-Leucine	AA	26.59	35.47	5.00	9.49	0.04	0.53	0.02	1.11	2.46
L-Methionine	AA	2.00	2.42	0.45	0.64	0.39	0.32	0.26	0.71	1.50
L-Phenylalanine	AA	4.59	5.45	0.90	1.69	0.61	0.21	0.42	0.52	1.25
L-Proline	AA	52.51	66.35	14.53	27.23	0.57	0.24	0.43	0.57	1.31
L-Tryptophan	AA	3.79	5.09	1.41	1.29	0.12	0.44	0.26	0.73	2.05
L-Tyrosine	AA	3.08	4.33	1.10	2.17	0.46	0.28	0.33	0.59	1.40
L-Valine	AA	46.83	61.89	10.18	16.76	0.10	0.47	0.04	1.05	2.31
Benzene acetic acid	AA	0.59	0.29	0.48	0.27	0.39	0.34	0.33	0.67	2.15
L-Aspartic acid	AA	0.44	0.32	0.22	0.23	0.57	0.27	0.57	0.51	1.37
Decanoic acid	FA	0.31	0.44	0.19	0.21	0.39	0.32	0.35	0.69	1.71
Dodecanoic acid	FA	0.47	0.69	0.21	0.45	0.57	0.23	0.35	0.52	1.59
Glycerol	FA	12.78	19.11	5.49	8.37	0.12	0.42	0.20	0.92	2.28
Nonanoic acid	FA	1.59	1.17	0.73	0.34	0.57	0.24	0.43	0.55	2.63
Glyceric acid	CA	2.34	2.80	0.73	0.69	0.40	0.31	0.39	0.64	1.32
Ribitol	CA	1.91	2.20	0.63	0.66	0.43	0.30	0.61	0.50	1.02
Unknown 1	U	1.06	1.49	0.66	1.08	0.57	0.23	0.57	0.56	1.26
Unknown 12	U	0.37	0.45	0.14	0.16	0.57	0.23	0.57	0.53	1.19
Unknown 13	U	0.56	0.73	0.34	0.29	0.61	0.20	0.57	0.51	1.13
Unknown 14	U	1.64	5.98	1.96	12.21	0.39	0.33	0.26	0.60	1.81
Unknown 2	U	0.22	0.32	0.12	0.15	0.40	0.31	0.33	0.64	1.58
Unknown 3	U	1.47	3.25	3.32	4.25	0.22	0.39	0.26	0.72	1.90
Uridine	U	1.09	1.39	0.42	0.46	0.40	0.31	0.43	0.60	2.05
Methanol	U	90.22	65.76	33.06	46.26	0.12	0.42	0.20	0.76	1.97

The Kegg IDs of the 24 significant metabolites in sub-study 1 were subject to pathway analysis in MetaboAnalyst (Chong *et al.*, 2018). Figure 5-12 shows the pathway impact and -log(p-values) for the specific pathway for this list of metabolites. Pathways identified to be affected include the valine, leucine and isoleucine biosynthesis pathway (p = 7.09E-05 ; impact =0), the phenylalanine metabolism pathway (p = 6.79E-03; impact = 0.35) and the phenylalanine, tyrosine and tryptophan biosynthesis pathway (p = 2.17E-02; impact = 1).

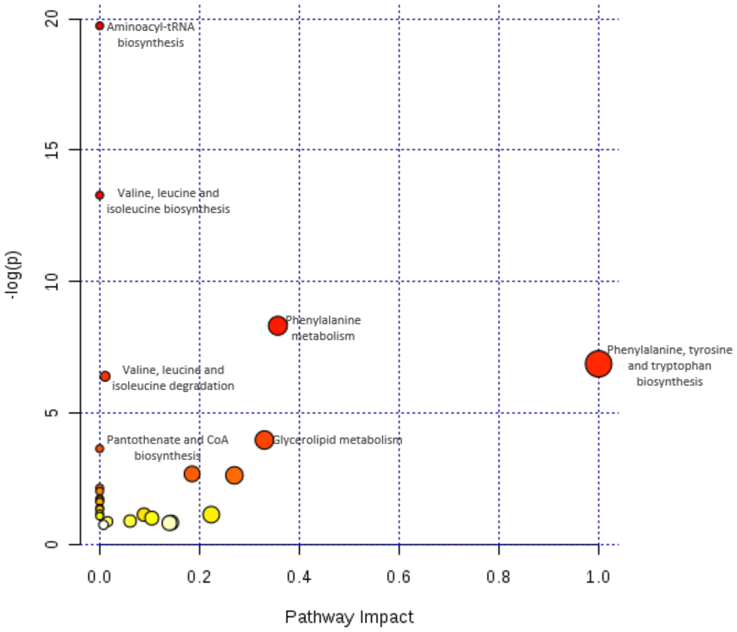

Figure 5-12: **Pathway analysis of the significant metabolites in CD4-high versus CD4-low groups.**

HIV/AIDS is characterised by a catabolic state, utilising protein and fats as the primary fuel source. Our results are in line with previous literature suggesting that the primary energy metabolism shifts towards amino acids, then fatty acids, and then carbohydrates (Aounallah *et al.*, 2016; Fitzpatrick & Young, 2013; Graber, 2001; Sitole *et al.*, 2013). Elevated proteolysis is expected to result in elevated concentrations of amino acids, as is known to occur during brief fasting or the initial phase of starvation (Pelley, 2012). This increase in amino acids is seen in the HLA non-protective HIV positive group in sub-study 3. This state of proteolysis significantly decreases during the later stages of starvation as a means to protect essential proteins during the wasting state. This reduction in proteolysis is what we anticipate is happening in the HIV positive patients with low CD4 counts, investigated in sub-study 1. This is seen by the significant decrease in 14 amino acids and their derivatives compared to the three decreased fatty acids and two decreased carbohydrates. Furthermore, chronic HIV infection is associated with increased resting energy expenditure and increased fatty acid oxidation (Hommes *et al.*, 1991). Increased resting energy expenditure, inadequate nutrient intake and malabsorption due to disturbed gut microbiome and mucosa, contribute to the malnutrition typically seen in HIV positive individuals (Hsu *et al.*, 2005). Besides the decreased absorption of sugars, which predisposes HIV positive individuals to lose weight, AIDS has been previously associated with cachexia [ongoing loss of skeletal muscle mass (Fearon *et al.*, 2011)] also called wasting syndrome (Fitzpatrick & Young, 2013; Salas-Salvado & Garcia-Lorda, 2001). Increased protein breakdown does not necessarily lead to higher concentrations of plasma amino acids, since decreased plasma amino acids were previously associated with HIV positive individuals (Fuchs *et al.*, 1990; Zangerle *et al.*, 2010). Metabolic reprogramming may be the driver behind the reduced amino acids detected, which eventually leads to muscle breakdown, a feature typically seen when comparing plasma of starving vs fasting individuals (Worthley *et al.*, 1984).

Branched-chain amino acids (BCAAs) are part of the essential amino acid group not synthesized in humans. BCAAs have been associated with innate immune function (Ma & Ma, 2019). BCAAs are associated with the regulation of protein synthesis and breakdown in skeletal muscle (Norton & Layman, 2006). Protein breakdown, as well as the type and quantity of various food sources, influence the BCAAs' plasma concentration, as depicted in Figure 5-13. The BCAAs leucine, isoleucine and valine, as well as the catabolic product of branched-chain amino acid transaminase 1 (BCAT1), 2-ketoisocaproic acid, were all decreased in the CD4-low group. Muscle wasting increases the availability of BCAAs. However, the concentration of BCAAs depends on the balance between protein breakdown and the catabolism of the BCAAs (Herman *et al.*, 2010; Holeček, 2018). Previous literature suggests mitochondrial damage in HIV positive individuals to additionally contribute to the increased BCAAs and their keto acid derivatives typically seen in these patients. Participants with lower CD4 counts are expected to be at a more progressed stage of HIV infection and would thus have more mitochondrial damage as well as higher levels of BCAAs and their keto acid derivatives, which is not the case in our study.

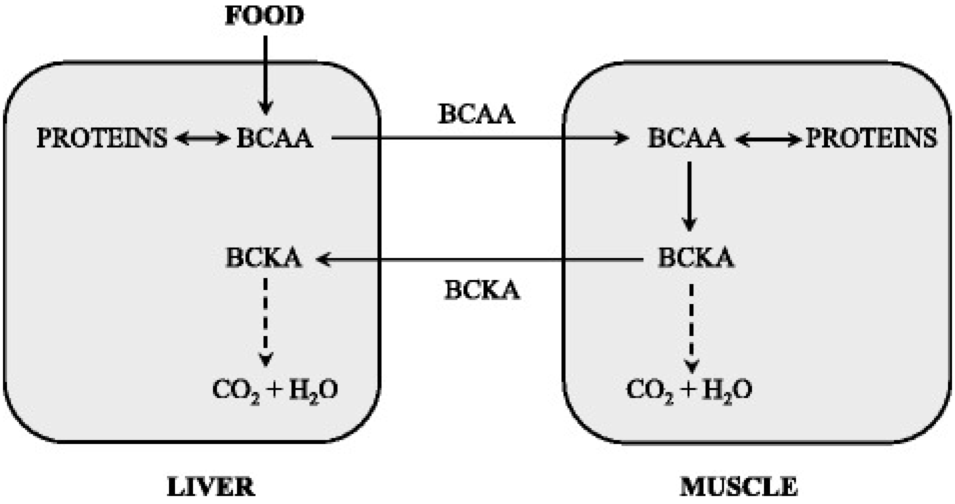

Figure 5-13: BCAA homeostasis from food and proteins. Figure from Holeček (2018) (used under fair dealing rights as described in the SA Copyright act)

A possible explanation for this is that increased inflammation results in a lack of appetite in the low CD4 count participants, thereby decreasing the amount of BCAAs from nutritional sources. This trend in BCAAs can similarly be attributed to malabsorption of amino acids in general, due to disturbances in the microbiome and gut mucosa typically seen in AIDS (Cassol *et al.*, 2013; Herrera *et al.*, 2019; Vesterbacka *et al.*, 2017). Another hypothesis is that HIV infection is associated with higher basal metabolic rates (Hommes *et al.*, 1991; Vassimon *et al.*, 2012) which require an increase in adenosine triphosphate (ATP) production. The latter is also synonymous with cachexia. Activation of adenosine monophosphate-activated protein kinase (AMPK) by adenosine monophosphate (AMP) is a key regulator of mitochondrial biogenesis, which initiates the proliferation of mitochondria (Jornayvaz & Shulman, 2010; Korzeniewski, 2001). An increase in mitochondrial function would increase the capacity for ATP production. This will deplete energy sources in order to supply the increased basal metabolic rate. Hence, increased oxidative phosphorylation could explain the decreased BCAAs in the CD4-low group.

Furthermore, the CD4-low group had decreased phenylalanine, tyrosine and tryptophan concentrations. BCAAs, phenylalanine and tyrosine share the same cell membrane transporters in the brain (Monirujjaman & Ferdouse, 2014). Similar trends in the BCAAs and these aromatic amino acids could indicate that brain amino acid metabolism likely plays a role in the reduction of both these classes of metabolites (supported by serotonin and dopamine levels, also detected here). Activation of the IDO-mediated tryptophan metabolism has previously been associated with HIV-induced immune activation, through microbial metabolism, due to dysregulation of the gut mucosal membrane (Herrera *et al.*, 2019). These aromatic amino acids, phenylalanine, and tyrosine, function as precursors for catecholamine neurotransmitters (dopamine), while tryptophan serves as a precursor for the monoamine neurotransmitters (serotonin) (Fernstrom & Fernstrom, 2007). Serotonin and dopamine are linked to neuro-conditions reported in HIV positive individuals (Cassol *et al.*, 2014; Gostner *et al.*, 2015a; Pendyala & Fox, 2010). The HIV-infected individuals with low CD4 counts, generally presented with decreased levels of these neurotransmitter precursor amino acids, which have been previously associated with the neuropsychiatric symptoms observed in HIV-infected individuals (Gostner *et al.*, 2015a). These reduced amounts of amino acids may be indicative of elevated serotonin and dopamine which have been previously associated with cachexia, a condition ascribed to the participants of sub-study 3. Decreased levels of hydroxyproline is associated with a deficiency in vitamin C, which has been observed in various HIV/TB co-infected cohorts (Semba *et al.*, 2010). Furthermore, decreased levels of hydroxyproline suggests decreased proteolysis (Wishart *et al.*, 2017) complementary to the starvation-like metabolite profile of this cohort. Given the severity of disease in the majority of this cohort, which is usually accompanied by decreased appetite and malabsorption (Gostner *et al.*, 2015b) as well as their low-income background, vitamin deficiencies are probable.

HIV-induced metabolic reprogramming is known to alter glucose metabolism. HIV infection is particularly associated with an increase in the expression of glucose receptors and the levels of glycolytic intermediates suggestive of elevated glucose metabolism (Ahmed *et al.*, 2018). It is speculated that carbohydrates are taken up by host cells and utilised by HIV for its replication, in turn enhancing immune activation, ROS production, inflammatory processes and mitochondrial damage (Sitole *et al.*, 2013). In a study comparing the metabolites of oral wash samples from HIV-infected individuals, Ghannoum *et al.* (2013) showed all carbohydrate compounds to be decreased in the samples of the infected patients. This supports the changes observed for this particular cohort. Reduced plasma carbohydrates in the HIV positive participants with low CD4 counts, could similarly to the reduced amino acid concentrations, be attributed to the aforementioned appetite suppression, malabsorption, and/or increased catabolism, which is well established in HIV infection (Lake & Currier, 2013; Palmer *et al.*, 2016; Sitole *et al.*, 2013), and synonymous with cachexia.

Odd chain fatty acids are associated more specifically with microbial metabolism (Wishart *et al.*, 2017), which is generally known to be dysregulated in HIV-infected individuals (Vázquez-Castellanos *et al.*, 2018). Benzene acetic acid, which is also detected in elevated concentrations in the HIV positive participants with comparatively lower CD4 counts, is also associated with microbial dysbiosis (Wishart *et al.*, 2017).

Uridine is a pyrimidine-derived nucleoside with the base uracil, occurring only in RNA. Uridine is incorporated into viral RNA during viral replication. The clinical data shows that the average viral load of the samples in the CD-low group is almost three times more than that of the CD4-high group and is inversely proportional to the uridine concentration. The reduced uridine in the CD4-low group may be due to its incorporation into viral RNA. However, the increase in the uridine precursor aspartic acid and decrease in uridine levels suggests a decrease in pyrimidine synthesis by the host. This is likely caused by the metabolic reprogramming focussed on providing cells with their energy requirements (Hommes et al., 1991).

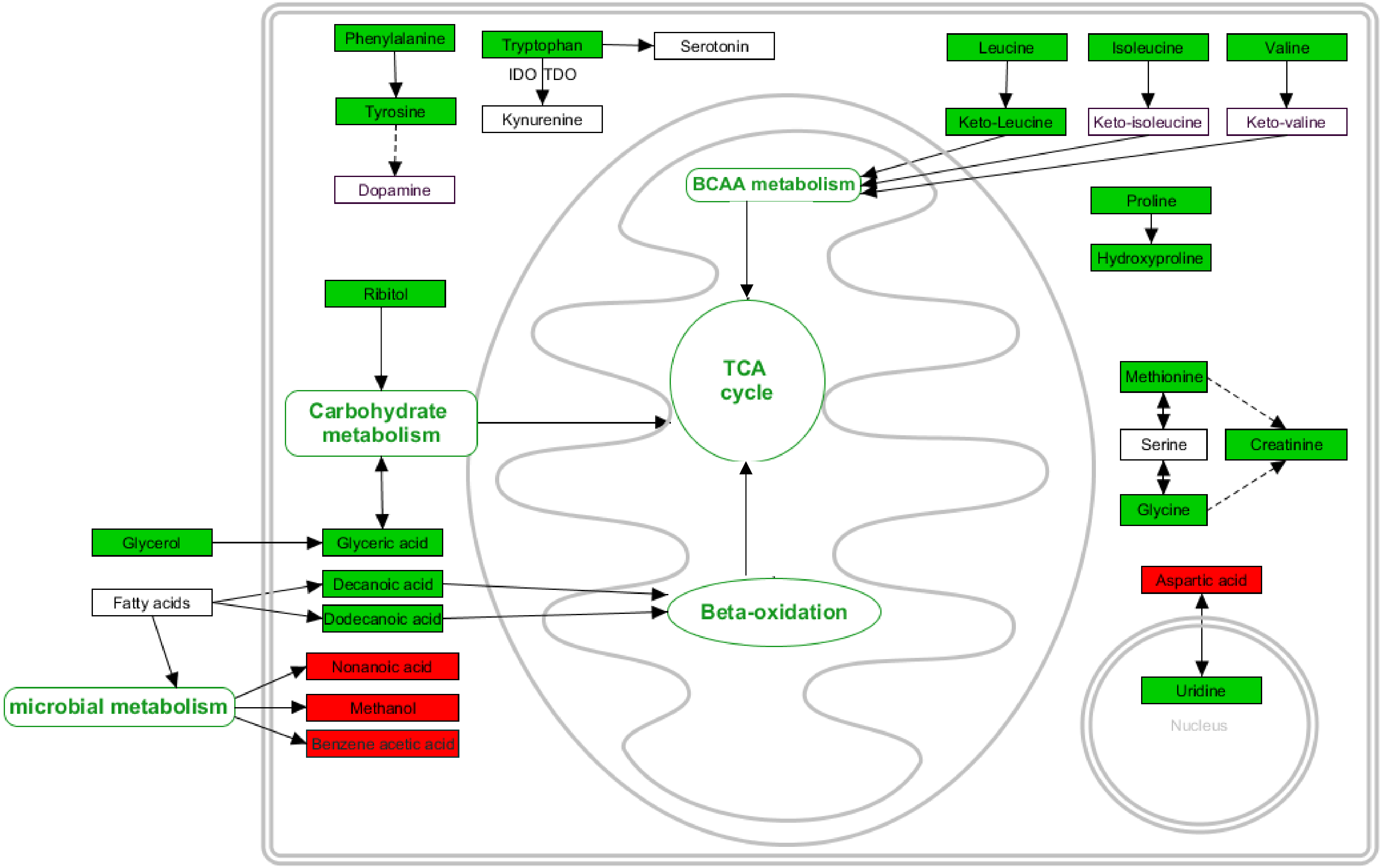

Figure 5-14 Metabolic pathways associated with the significantly altered metabolites in the HIV positive group with low CD4 counts. Metabolites shown in green and red are decreased and increased, respectively in the CD4-low group. Metabolites in black are associated but were not detected. IDO:Indoleamine Deoxygenase, TDO: Tyrosine deoxygenase, TCA: Tricarboxylic acid

.

5.5.2 Non-progressors versus progressors

Table 5-4 lists the metabolites deemed significant between HIV positive individuals classified as non-progressors and progressors based on the change in CD4 over time. The colours of the means in the table show the increased value in red and the decreased value in green. Statistically significant (MW adjusted p-value ≤ 0.05; MW effect size ≥ 0.5; Unpaired t-test adjusted p-value ≤ 0.05; ES Cohen’s d-value ≥ 0.8) and slightly significant (MW adjusted p-value ≤ 0.1; MW effect size ≥ 0.3; Unpaired t-test adjusted p-value ≤ 0.1; ES Cohen’s d-value ≥ 0.5) metabolites between the non-progressor and progressor groups are coloured in lime and yellow, respectively. The PLS-DA variable importance on projection is also included, but not used for classification of the metabolites in this instance. The data is merely shown to illustrate that the multivariate analysis complements those metabolites found to be of significance through univariate tests. When considering the change in CD4 over time, progressors reflect an increase in certain unresolved carbohydrates with slight significance.

Table 5-4: Table of statistically significant metabolites measured between the samples of non-progressors and progressors (CA = carbohydrates, U = Unknown). Elevated metabolites are highlighted in red, and those reduced in green.

		Untransformed						Transformed		
		Mean (mmol/l)		Standard deviation		Mann-Whitney test		t-test	Effect size	
Metabolite	Class	Non-progressor	Progressor	Non-progressor	Progressor	Adjusted P-value*	Effect size	Adjusted p-value*	Cohen's d-value	PLS-DA VIP**
Carbohydrate 2	CA	8.18	22.59	20.504	39.564	1.000	0.291	0.987	0.555	2.308
Carbohydrate 3	CA	2.34	5.04	2.840	4.272	1.000	0.295	0.987	0.657	2.768
Unknown 11	U	2.97	2.68	0.476	0.563	1.000	0.233	0.987	0.526	2.137
Unknown 14	U	5.22	2.70	6.787	2.440	1.000	0.229	0.987	0.500	2.126

Pathway analysis could not be performed due to the small number of significant metabolites and their uncertain identities. Statistically, this classification of HIV disease progression contributes the least to the metabolic differences detected considering the sub-studies.

All that can really be discussed about this group is the increase ~~decrease~~ in carbohydrates detected in the plasma of the HIV positive progressors (selected on the basis of change of CD4 counts over time), which may be associated with the aforementioned cachexia (Section 5.5.1). Changes in carbohydrate metabolism is a common feature observed in HIV positive patients. HIV-induced metabolic reprogramming is known to alter glucose metabolism. During infection where HIV is actively replicating, there is generally an increase in the expression of glucose receptors and the levels of glycolytic intermediates suggestive of elevated glucose metabolism (Ahmed *et al.*, 2018). Because glucose is constantly metabolised, its actual levels/concentration decreases. Since the analysed samples are at similar CD4 count although defined as fast and slow progressors based on the change in CD4 over time, increased carbohydrates in the progressive group suggests that because the CD4 counts are still generally mid-range and the individuals relatively "healthy", there may be less active viral replication taking place thus less utilization of glycolysis meaning less glucose is used and its concentrations actually increase. Differences in the metabolic profiles of actively replicating cells vs those experiencing latency supports this explanation (Castelano *et al.,*2019). GC-MS based techniques, however, have inherent difficulties in identifying sugars leaving the findings open for further research.

Comparing the results of sub-study 2 with that of sub-study 1, it is evident that the variation in CD4 counts in untreated HIV positive participants, has a much more pronounced effect on the metabolism, than the change in CD4 counts over time at similar CD4 counts (selected specifically for this reason as to eliminate the effect of CD4 counts in HIV disease progression in this study). In the study by Scarpelini *et al.* (2016), several metabolites including alpha amino adipic acid, lysine and branched chain amino acids were identified to distinguish rapid progressors from elite controllers in an HIV population. Upon investigating the demographics and clinical data of the compared cohorts in the study, it was found that the average CD4 count of the groups also differed significantly, confirming the role of CD4 counts in the altered metabolic state of HIV progressors. It therefore makes sense that the metabolic profile of the progressor cohort used and the trend of the molecules detected in their study is like that measured in our CD4-low cohort.

5.5.3 Protective versus non-protective *HLA-B* alleles

Table 5-5 lists the metabolites deemed significant between HIV positive individuals with protective and non-protective *HLA-B* alleles. The colours of the means in the figure show the increased value in red and the decreased value in green. Statistically significant (MW adjusted p-value ≤ 0.05; MW effect size ≥ 0.5; Unpaired t-test adjusted p-value ≤ 0.05; ES Cohen's d-value ≥ 0.8) and slightly significant (MW adjusted p-value ≤ 0.1; MW effect size ≥ 0.3; Unpaired t-test adjusted p-value ≤ 0.1; ES Cohen's d-value ≥ 0.5) metabolites between individuals with protective and non-protective *HLA-B* alleles are coloured in lime and yellow, respectively. The PLS-DA variable importance on projection is also included, but not used for classification of the metabolites in this instance. The data is merely shown to illustrate that the multivariate analysis complements those metabolites found to be of significance through univariate tests. Samples grouped according to

HLA-B alleles presented with more of a mixture in terms of class of compounds identified. In this group, amino acids, fatty acids and some sugars differentiated the metabolite profiles of individuals with protective and non-protective *HLA-B* alleles. The generalised trend was an increase in these metabolites in samples from participants with non-protective *HLA-B* alleles. Also notable is the very high levels of urea in the group with non-protective *HLA-B* alleles.

Table 5-5: Table of statistically significant metabolites measured between samples of participants with non-protective and protective *HLA-B* alleles (AA = Amino acids, FA = Fatty acids, CA = carbohydrates, U = Unknown). Elevated metabolites are highlighted in red, and those reduced in green.

		Untransformed						Transformed		
		Mean (mmol/l)		Standard deviation		Mann-Whitney test		t-test	Effect size	
Metabolite	class	Non-protective	Protective	Non-protective	Protective	Adjusted P-value*	Effect size	Adjusted p-value*	Cohen's d-value	PLS-DA VIP**
L-Isoleucine 2	AA	0.90	1.26	0.39	0.32	0.26	0.40	0.09	0.85	2.01
Oxalic acid	AA	1.20	1.63	0.48	0.80	0.65	0.25	0.59	0.53	1.91
Serine	AA	0.04	0.07	0.06	0.05	0.26	0.40	0.59	0.52	1.45
carbamate	AA	2.66	0.82	3.16	0.64	0.25	0.44	0.09	0.84	1.97
Creatinine	AA	11.02	7.60	6.64	4.34	0.62	0.30	0.55	0.55	1.16
Glycine 1	AA	7.91	5.70	4.96	3.81	0.65	0.23	0.65	0.52	1.32
Glycine 2	AA	37.01	25.89	16.07	7.88	0.61	0.33	0.34	0.79	1.62
Glycine 3	AA	1.26	0.74	0.82	0.30	0.25	0.43	0.09	0.88	2.26
L-Lysine	AA	0.94	0.59	0.55	0.28	0.62	0.29	0.34	0.67	1.61
L-Methionine	AA	2.48	2.04	0.86	0.57	0.65	0.24	0.55	0.52	1.28
Ornithine	AA	1.07	0.74	0.62	0.41	0.65	0.25	0.54	0.58	1.33
Pipecolic acid	AA	0.72	0.57	0.24	0.15	0.63	0.29	0.42	0.61	1.36
Serotonin	AA	0.38	0.20	0.33	0.21	0.61	0.32	0.48	0.59	1.38
Urea 1	AA	813.31	638.90	325.46	147.40	0.69	0.20	0.49	0.55	1.34
Urea 2	AA	862.48	250.43	1967.95	717.52	0.61	0.31	0.42	0.69	1.39
Carbohydrate 3	CA	2.90	5.67	3.18	4.71	0.65	0.25	0.59	0.53	2.06
Arabinoic acid	CA	0.24	0.14	0.15	0.08	0.62	0.29	0.20	0.68	1.57
Carbohydrate 1	CA	16.41	12.21	28.48	22.68	0.61	0.32	0.88	0.34	0.77
Glycerol monostearate	FA	2.74	2.12	0.99	0.69	0.63	0.29	0.48	0.61	1.42
Heptanoic acid	FA	1.41	1.23	0.18	0.25	0.31	0.38	0.48	0.67	2.38
Octadecanoic acid	FA	3.21	2.70	0.84	0.71	0.71	0.18	0.55	0.64	1.44
Palmitelaidic acid	FA	1.81	1.11	1.57	0.96	0.65	0.22	0.59	0.52	1.34
Glycolic acid	U	3.40	3.91	1.14	0.80	0.65	0.23	0.54	0.51	1.45
Unknown 7	U	2.91	3.49	0.78	0.76	0.61	0.32	0.44	0.69	1.41
2-Aminomalonic acid	U	7.16	4.23	3.72	1.58	0.26	0.42	0.09	0.94	2.23
Paracetamol	U	21.35	19.05	2.93	1.59	0.25	0.47	0.09	0.84	1.97
Unknown 10	U	3.51	3.00	0.98	0.82	0.65	0.25	0.59	0.54	1.30
Unknown 12	U	0.51	0.33	0.18	0.10	0.05	0.56	0.04	1.12	2.38
Unknown 13	U	0.72	0.50	0.32	0.38	0.61	0.33	0.54	0.62	1.44
Unknown 4	U	4.72	3.82	1.55	0.55	0.65	0.25	0.26	0.59	1.52
Unknown 5	U	7.35	6.47	1.52	0.73	0.65	0.21	0.40	0.56	1.46
Unknown 6	U	5.07	3.84	1.86	1.61	0.65	0.26	0.55	0.57	1.96
Unknown 8	U	2.07	0.40	3.16	0.25	0.69	0.20	0.20	0.55	1.25
Unknown 9	U	2.80	2.13	1.72	0.67	0.65	0.24	0.54	0.54	1.60

The significantly altered metabolites from sub-study 3 were subject to pathway analysis in MetaboAnalyst based on their Kegg IDs (Chong *et al.*, 2018). Figure 5-15 shows the pathway impact and -log(p-values) for the specific pathway for this list of metabolites. Pathways identified include the arginine biosynthesis pathway (adjusted-p =0.26; impact = 0.06), the glyoxylate and dicarboxylate metabolism pathway (adjusted-p = 0.18; impact =0.22) as well as the glycine, serine and threonine metabolism pathway (adjusted-p =0.58; impact = 0.46). However, the predictive p-values for these pathways suggest that they are of low significance.

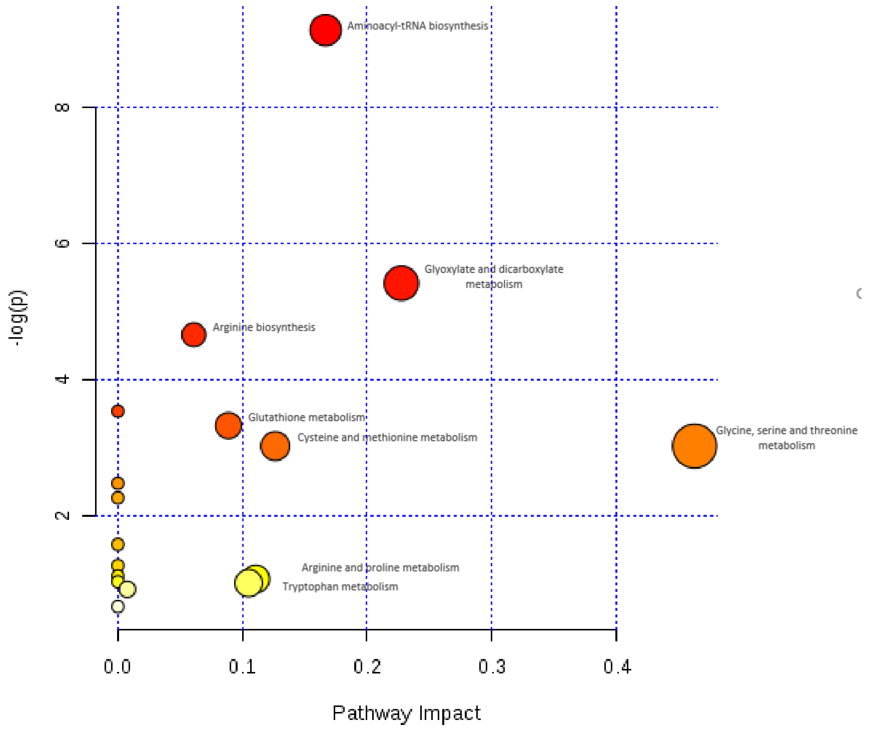

Figure 5-15: Pathway analysis of significantly different metabolites between the samples of participants with protective and non-protective *HLA-B* alleles.

Unknown compound 12, is a novel metabolite, significantly elevated in the plasma of the untreated HIV positive group with non-protective *HLA-B* alleles, considering all the statistical measures used. Unfortunately, however, we were unable to identify this compound using the NIST library. This unknown compound is possibly of great interest as it could be a potential biomarker of *HLA-B* predictive HIV disease progression. (Table 5-5). Figure 5-16, shows the three-dimensional total ion chromatogram (TIC) of the area surrounding the unknown peak. The highlighted area (marked x) in Figure 5-16 shows the unknown peak. The peak is so small that it is not visible on the TIC.

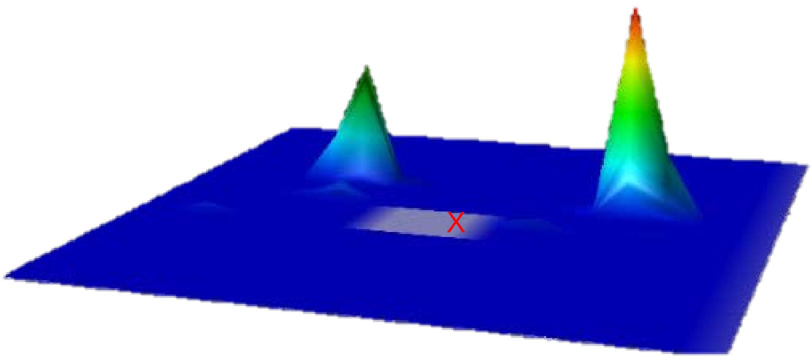

Figure 5-16: 3D-TIC highlighting the unknown compound, which was significant in all four statistical tests when comparing the samples of participants with protective and non-protective *HLA-B* alleles, respectively.

This peak was also inspected on the 2D TIC where it became clear that it was part of two co-eluting peaks that had to be deconvoluted. Figure 5-17 shows the chromatogram of the co-eluting peaks. The 73 m/z trimethyl silane ion ($C_3H_9Si^-$) is shared across derivatised molecules. Both the 192-B and 193 peaks in Figure 5-17 have apexes of the ion. Figure 5-18 shows the deconvoluted spectra of the unknown peak, peak 192-B in Figure 5-17. The 73 and 131 m/z ions both apex at the same retention time (RT) in the chromatogram in Figure 5-17. Fortunately, the unique mass (in this case 131 for peak 192-B) was used to quantify the compound. This metabolite, although unknown, has been sufficiently deconvoluted and differs statistically significant between samples from participants presenting different *HLA-B* alleles. Although the absolute identification of this peak would have been advantageous, it falls beyond the scope of this MSc investigation, since such a procedure is extremely challenging and time-consuming.

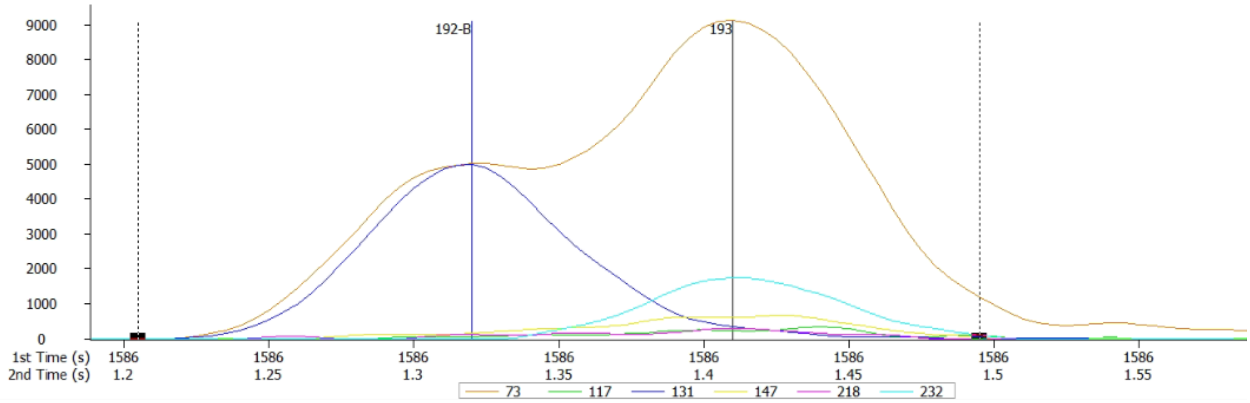

Figure 5-17: Deconvolution of the unknown peak (192-B) found to be significant across all statistical tests when comparing the samples of participants with protective and non-protective *HLA-B* alleles.

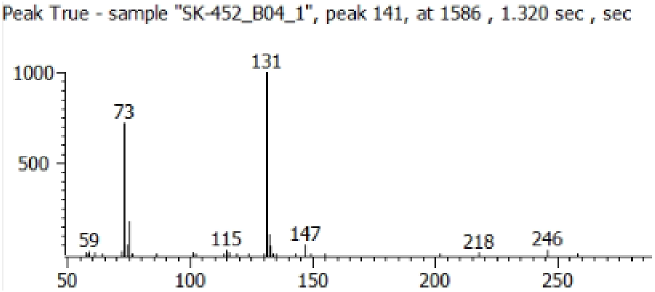

Figure 5-18: Deconvoluted mass spectra of the unknown peak found to be significant across all statistical tests when comparing the samples of participants with protective and non-protective *HLA-B* alleles.

A general increase of amino acids, fatty acids and carbohydrates can be seen in the HIV positive individuals with non-protective *HLA-B* alleles compared to those with the protective alleles. The increase in amino acids is likely due to an earlier activation of the wasting state. Wasting in this group can be seen as the increase in amino acids driven by proteolysis. This is still possible in these individuals, due to the fact that they were selected from a cohort with CD4 counts of between 350-499 cells/µl, hence, not as progressed as the HIV patients in the CD4-low group in sub-study 1. The participants of sub-study 3 can be compared to individuals experiencing a brief fasting state, or the initial phase of starvation, where wasting has not yet occurred, and proteolysis is a viable option for energy production. A previous study by Worthley *et al.* (1984) on patients undergoing a period of brief fasting, showed similar alterations to their plasma amino acid metabolomes (reduced leucine, isoleucine, valine, lysine and serine and elevated glycine, phenylalanine, methionine and alanine), as we did in the *HLA-B* non-protective HIV group of sub-study 3, confirming our hypothesis. Increased ornithine and urea indicate an up-regulation of the urea cycle as a means to excrete excess ammonia from the increased catabolism of the amino acids (Conway & Hutson, 2016; Holecek, 2002). Creatinine concentration is also directly related to the urea cycle, especially ornithine (Schutte *et al.*, 1981).

Increased C16 and C18 fatty acids as measured here have previously been associated with late-stage HIV-1 replication (Kulkarni *et al.*, 2017). Octadecanoic acid, also called stearic acid, is formed from the dissociation of glycerol from glycerol monostearate.

Both arabinoic acid and heptanoic acid are of microbial origin and are increased in the group with non-protective *HLA-B* alleles. Evidence suggests that pipecolic acid is a microbial product of lysine metabolism (Broquist, 1991). The elevated lysine is expected from proteolysis which may be driving the increase in pipecolic acid. This suggests that the non-protective *HLA-B* allele group is associated with an increased dysbiosis of the gut microbiota as previously shown in the literature (Vujkovic-Cvijin *et al.*, 2013). Vujkovic-Cvijin *et al.* (2013) also linked tryptophan catabolism to disease progression. Our results are in line herewith showing an increase in serotonin, a catabolic product of tryptophan. Oxalic acid and glycolic acid are related, and both were included in the glyoxylate and dicarboxylate pathway which was prominent in the pathway analysis. Glyoxylate is an intermediate of the glyoxylate cycle which enable plants and microbes to convert fatty acids into carbohydrates. The acid form, glycolic acid, was decreased here, confirming the use of the intermediate for sugar production. This pathway is, however not known to occur in mammals (Traven & Naderer, 2019). Oxalic acid and glycolic acid are also of microbial origin. The decrease in the glyoxylate pathway products along with an increase in an odd chain fatty acid, the breakdown of lysine to pipecolic acid and increased arabinoic acid indicates that microbial metabolism differs between HIV-infected individuals with non-protective and protective *HLA-B* alleles. That *HLA* alleles would be associated with alterations to the gut microbiota is novel yet not surprising given the role of these proteins in presenting antigens to immune cells (Borghans *et al.*, 2007). Ineffective presentation of antigens by *HLA-B* would for example result in increased viral turnover, increased immune activation and a leaky gut mucosa. Some carbohydrates such as arabinoic acid increased and others such as “carbohydrate 3” decreased and may possibly inform on specific metabolic changes attributed to gut microbes versus HIV-induced host metabolic changes.

Aminomalonic acid was also significantly increased in the non-protective *HLA-B* group. No biologically relevant metabolic pathways have been linked with aminomalonic acid. Possible origins of aminomalonic acid include oxidative damage to proteins and errors in protein synthesis (Wishart *et al.*, 2017). Structurally, aminomalonic acid is similar to serine. NAD$^+$ dependent serine 3-dehydrogenase is a microbial enzyme which converts serine to 2-aminomalonate semialdehyde. Therefore, microbial degradation of serine is likely attributed to decreased serine and increased 2-aminomalonic acid.

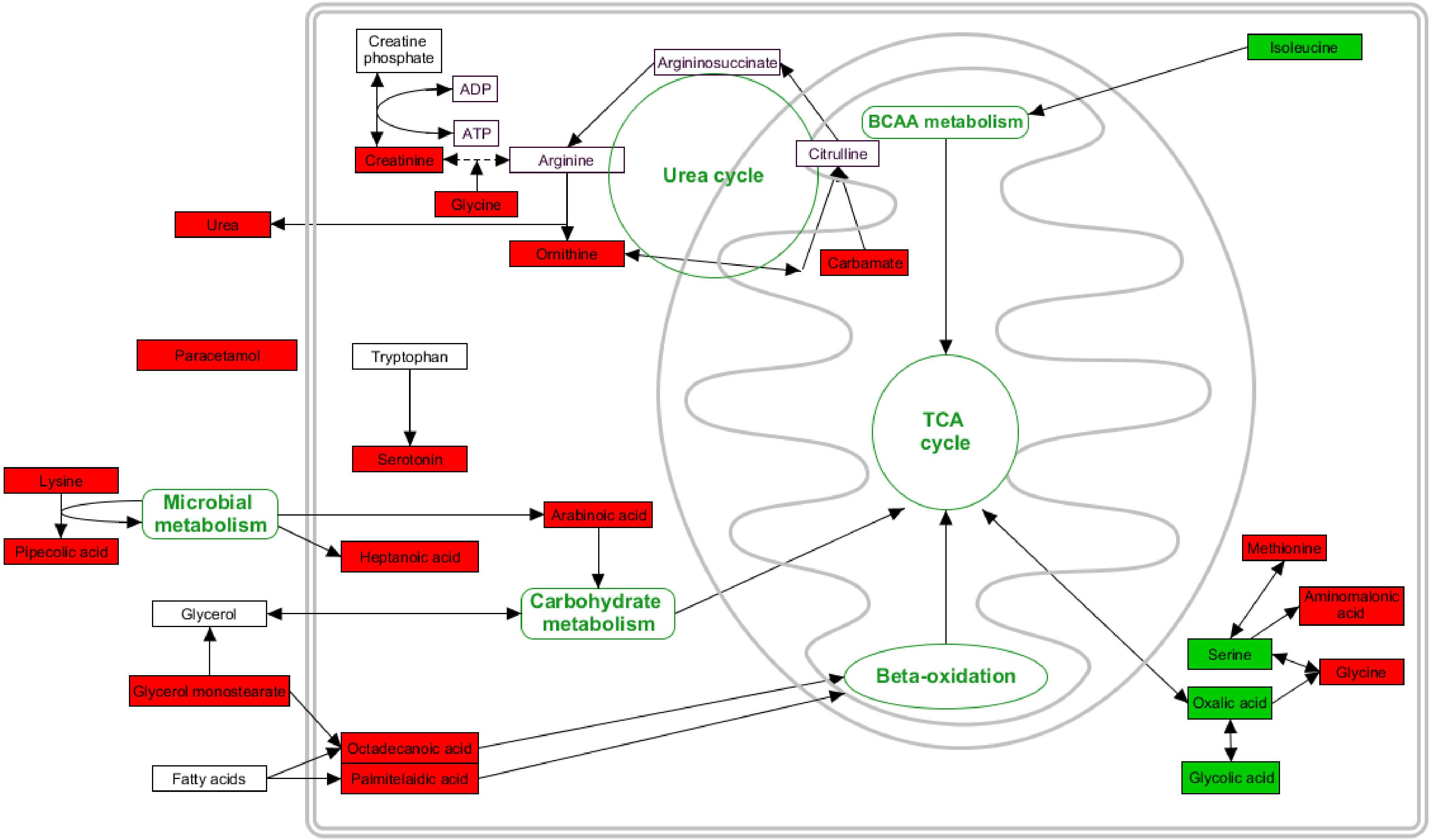

Figure 5-19: Metabolic pathways associated with the significantly altered metabolites in the HIV positive group with non-protective *HLA-B* alleles. Metabolites shown in green and red are decreased and increased, respectively in the non-protective *HLA-B* group. Metabolites in black are associated but were not detected.

5.6 Considerations on interpreting the findings

The findings reported in this study should be interpreted while considering the limitations of this study. The nature of this study involves treatment-naïve participants. A study on treatment-naïve individuals would not be considered ethical since the test and treat policy was adopted, therefore, previously collected samples had to be used. This introduced sample age as a factor. The quality of plasma stored at -80°C remains a concern despite freezer temperatures being monitored and access to the fridges being controlled. The storage of samples for periods greater than 10 years increases the chances of undetected fluctuating storage conditions which might influence metabolomics data. Although plasma metabolites were reported to be stable when stored at -80°C (Breier *et al.*, 2014; Hebels *et al.*, 2013), newer literature suggests otherwise (Stevens *et al.*, 2019). Although sample ages were > 10 years, this parameter was not statistically significant between the comparative groups but may have contributed to low intensity signals/loss of some analytes and their subsequent non-detection.

Additionally, only the demographic and clinical parameters collected with the samples could be used. This lead to CD4 being the primary parameter used in participant and sample selection, ignoring VL which was missing for some samples. An advantage of this however was that the selection criteria for sub-study 1 for example allowed for comparison of extremes, so allowing a greater chance of detecting differences.

We reported the participant and time-point selection criteria in Sections 3.4 and 3.5 respectively. Data interpretation is thus specific to the selected samples since the previously collected samples were not accompanied with dates of primary infection thereby preventing us from determining the stage of infection. To distinguish between all the definitions of progression, at least ten years' follow up data is needed. For sub-study 2, a minimum of four years' follow up data was used. The definition of progression as used in sub-study 2 was thus based on the data available to us while still considering definitions in literature. Although four years' worth of CD4 measurement records had to accompany the samples for assessment of clinical progression, four years of follow up may be insufficient, or the CD4 decline may not be sufficiently different over this time to allow for detecting differences associated with progression hence the limited alteration to the metabolome as measured for sub-study 2.

With regard to the *HLA-B* alleles, not everyone with a protective *HLA-B* allele will necessarily be a non-progressor and not everyone without a protective *HLA-B* allele will necessarily be a progressor. For example, among HIV LTSP/LNTPs, approximately 50% might have a protective allele.

Furthermore, peak identities were assigned through spectral matching constituting a level 3 identification. With a level 3 identification many compounds/analytes cannot be distinguished, especially carbohydrates, leaving room for the discovery of potential biomarkers that are significant yet currently unidentifiable. Considering all these limitations the experimental design and the execution thereof was done in the most optimal way to produce biologically relevant data which could be defended/supported by existing literature While the findings obtained here shed light on mechanisms impacting progression, AIDS remains a complex disease and likely involves multiple factors contributing to a particular phenotype.

FINAL CONCLUSIONS

6.1 Concluding summary

HIV/AIDS has received a lot of media attention in previous years. The treatment of HIV-infected individuals has prolonged patient life expectancy, which has subsequently led people to become complacent with the idea of HIV infection (Abdool Karim & Abdool Karim, 2018; Hurley, 2018). HIVAIDS, however, remains a significant health concern worldwide.

In the absence of HAART, statistics generally report an average of eight years for progression from HIV infection to AIDS (Scarpelini et al., 2016). There are however unique HIV subpopulations that do not follow this generalised survival trend with some individuals progressing faster in the disease than others. Those populations with slower progression rates have received much attention in physiological, genetic and proteomic studies, but the mechanisms of slower progression remain unclear.

The assessment of HIV status, disease progression (Hsu *et al.*, 2005) and monitoring of HIV treatment (Palmer *et al.*, 2016), are currently exclusively based on patient CD4 counts. Unfortunately, CD4 counts are not exclusively determined by HIV infection. Thus, characterisation of disease progression needs novel biomarkers. Specific immunogenetic parameters have previously been associated with HIV disease progression. HIV/AIDS is additionally associated with an increase in basal metabolic rate (Salas-Salvado & Garcia-Lorda, 2001) and markers of mitochondrial damage (Masson *et al.*, 2017; Rodríguez-Gallego *et al.*, 2018b; Sitole *et al.*, 2013). The interaction between HIV and the metabolism shows promise for identifying new metabolic markers with which to better characterise HIV disease.

For the purpose of studying the altered metabolite profile of different clinical and immunogenetic factors associated with HIV disease progression, the UKZN agreed to collaborate, by sharing previously collected plasma samples with the required clinical and immunogenetic data we required. Previously standardised extraction and analysis methods were used to obtain the metabolite profile of these samples. The analysis methods passed several quality control assessments and could thus be applied to test samples. The analysed samples were then subjected to statistical analysis and biological interpretation.

In the current investigation, participants with low CD4 counts presented with an increased catabolic state associated with wasting and cachexia (characterised by elevated energy expenditure, loss of appetite and malabsorption), and the associated general decrease of plasma amino acids, fatty acids and carbohydrates. The number of significantly decreased metabolites in these metabolite classes suggest a shift in primary energy metabolism away from amino acids. The metabolite profiles of the individuals with low CD4 counts signified a state of long term starvation and microbial dysregulation.

HIV disease progression (characterised by a significant negative correlation between CD4 and time) induced far less of a change to the HIV positive metabolome. The only observable difference between those classified as progressors and non-progressors were two carbohydrates, which were increased in the HIV positive progressor group. The findings here were thus inconsistent with the study of Scarpelini *et al.* (2016) who showed five metabolites to be decreased and to distinguish rapid progressors at baseline. Their findings were mainly deregulation of acylcarnitine and sphingomyelin metabolism, down regulation of beta-oxidation and sphingosine-1-phosphate-phosphatase-1 activity and elevated levels of acyl-alkyl-containing phosphatidylcholines and alkylglyceronephosphate synthase in immunological non responders. CD4 count was, however, found to be a confounder in their design hence the measured metabolites reflect an opposite trend to our data.

HIV-infected individuals with non-protective *HLA* alleles, on the other hand, showed increased levels of plasma amino acids, carbohydrates, fatty acids and microbial metabolites. The increased amino acids and urea suggests elevated proteolysis and catabolism of proteins, typically seen during a state of early fasting, before wasting (Herman *et al.*, 2010). The fatty acids that were increased are associated with late-stage viral replication, indicating increased viral turnover (possibly as a result of ineffective presentation of antigens by the non-protective *HLA-B* alleles). A large number of metabolites were identified and indicative of dysregulation of the microbiome in the non-protective *HLA-B* group, comparatively.

Previous metabolomics studies on HIV disease progression (Scarpelini et al., 2016; Zanoni *et al.*, 2017) used much smaller cohorts (n=5 per group) which were mainly infected with HIV subtype B which is not representative of the predominant HIV subtype C strain found in South Africa. Although Tarancon-Diez *et al.* (2019) employed both untargeted and targeted GC and LCqTOF approaches to investigate mechanisms that differentiate individuals who spontaneously lose virological control over those who maintain it, the GCxGC-TOFMS system used in our study has still not been employed for progression-based analysis. Our work, therefore, expands on the work of Scarpelini, Zanoni and Tarancon-Diez by reporting metabolic differences in a larger South-African cohort representative of HIV subtype C infections. In addition to measuring metabolite profiles at low and high CD4 counts, this study also compares progressors defined by a change in CD4 over time and individuals presenting non- protective *HLA-B* alleles, at a median CD4 count thus reporting data where CD4 is not a confounder.

The trends observed suggest that CD4 count is associated with metabolic change. When CD4 count is not a confounder there is minimal metabolic variation between samples within a group. This makes sense since various other factors, not accounted for, may be contributing to the metabolic profile. When we investigate for example the metabolic profile of plasma from participants with non-protective *HLA-B* alleles, we see that the metabolic changes measured in sub-study 3 present a mix class of metabolites which are distinct from that of sub-study 1 yet it overlaps with the trends of sub-study 2 in that carbohydrates are generally increased. In sub-studies 2 and 3, CD4 as confounder has been removed supporting the similar trends observed in these sub-studies.

Nonetheless, this untargeted metabolomics approach provides new insight into the role of clinical and immunogenetic factors linked to HIV disease progression. Where a clinical parameter such as low CD4 count is used to metabolically define patient prognosis, protein catabolism mainly suggests use of these breakdown products for energy production. When immunogenetic parameters are used to metabolically define patient prognosis, the non-protective *HLA-B* alleles most likely due to their limited ability to present antigen and initiate effective immune responses, facilitates viral replication, increasing immune activation and in turn disrupting the gut microbiome. The dissimilarity in the trends of metabolite classes between sub-study 1 versus that of sub-study 2 and 3 indicates that rather than a single metabolic mechanism involved in the control of HIV, there are many pathways involved which are differentially modulated. Nonetheless, participant groups with a “poorer” outcome generally showed features with some similarity implying that a holistic view into differential HIV disease progression in these patients may benefit from the inclusion of a metabolic component.

6.2 Prospects in the metabolomics of HIV disease progression

From the results obtained in this study, the following analytical approaches on a similar cohort of participants presenting with differential HIV disease progression is advised:

- Source untreated cohort with more complete demographic and clinical data over longer time period , especially relevant to sub-study 2
- Targeted analysis of amino acids will increase the statistical power of such a metabolomics study and will give a clearer picture of the wasting state hypothesised at different CD4 counts and/or stages of disease progression.
- Targeted analysis of carbohydrates with better resolution of isomers will shed light on the principles of energy metabolism at different CD4 counts and/or stages of disease progression.

- Targeted analysis of fatty acids and acylcarnitines will assist in better understanding the altered fatty acid metabolism identified at different CD4 counts and/or stages of disease progression.
- Including samples from multiple cohorts to increase the number of samples, would inherently increase the statistical power of such an investigation.
- The identification of the unknown significant metabolite seen to be elevated in the non-protective *HLA-B* allele group may shed light on the mechanisms involved and/or serve as an early indicator of progression.
- Future work may entail further subdivision of the *HLA-B* alleles into protective and non-protective and characterising the metabolic profiles of these.
- Immunometabolism refers to the convergence of bioenergetic pathways and the specific functions of immune cells, as defined by (Shi *et al.*, 2016). Altered levels of cytokines have been studied in HIV infection (Williams, 2012), and are known to be linked to various metabolic changes, though these interactions are complex and not fully understood as yet (Sauerwein *et al.*, 2011). Given the interplay between the immune and metabolic systems we investigated in parallel to the metabolic data, changes in Th1/Th2/Th17 plasma cytokine levels for sub-study 1 through to 3. Preliminary results show increased IL-2, IL-4 and IL-6 levels in the progressor group defined in sub-study 2, although the metabolite profiles for this sub-study did not show much metabolic variation except for two carbohydrates. When we look at the cytokine profile for evaluating HIV disease progression, a shift from T-helper type 1 (Th1) cell-mediated immunity to T helper type 2 (Th2) humoral immunity is generally observed in cases experiencing disease progression. T-helper type 1 (Th1) cells produce (IL)- 2, and antiviral interferon (IFN)- γ while T helper type 2 (Th2) cells produce cytokines; IL-4, IL-10 and proinflammatory cytokines. IL-6. IL-2 and IL-4 regulate T-cell proliferation and survival. The maintenance of a viral reservoir to the detriment of the host is reflected by the increase in these cytokines. IL-6 is a proinflammatory cytokine and has previously been shown by (Williams, 2012) to be associated with progressive HIV infection. Covariate analysis between the cytokine data and the metabolite profiles thus stands to inform on the immuno-metabolic mechanisms defining differential disease progression. Information to this analysis is indicated in Appendix 1.
- In our findings, we noted metabolites of microbial, viral and host origin. To clarify the mechanisms clearly, an analysis such as this may benefit by designing experiments in such a way so as to clarify the origin of metabolites, perhaps by investigating the enantiomers of certain metabolites.

BIBLIOGRAPHY

Abdool Karim, Q. & Abdool Karim, S.S. 2018. HIV—No time for complacency. *Science*, 360(6394):1153-1153.

Ahmed, D., Roy, D. & Cassol, E. 2018. Examining relationships between metabolism and persistent inflammation in HIV patients on antiretroviral therapy. *Mediators of inflammation*, 2018.

Anthony Nolan Research Institute. 2019. HLA Alleles Numbers. http://hla.alleles.org/nomenclature/stats.html Date of access: 2019-07-14.

Aounallah, M., Dagenais-Lussier, X., El-Far, M., Mehraj, V., Jenabian, M.A., Routy, J.P. & van Grevenynghe, J. 2016. Current topics in HIV pathogenesis, part 2: Inflammation drives a Warburg-like effect on the metabolism of HIV-infected subjects. *Cytokine Growth Factor Reviews*, 28:1-10.

Armah, K.A., McGinnis, K., Baker, J., Gibert, C., Butt, A.A., Bryant, K.J., Goetz, M., Tracy, R., Oursler, K.K. & Rimland, D. 2012. HIV status, burden of comorbid disease, and biomarkers of inflammation, altered coagulation, and monocyte activation. *Clinical Infectious Diseases*, 55(1):126-136.

Bardeskar, N.S. & Mania-Pramanik, J. 2016. HIV and host immunogenetics: unraveling the role of HLA-C. *HLA*, 88(5):221-231.

Barnes, S., Benton, H.P., Casazza, K., Cooper, S.J., Cui, X., Du, X., Engler, J., Kabarowski, J.H., Li, S. & Pathmasiri, W. 2016. Training in metabolomics research. II. Processing and statistical analysis of metabolomics data, metabolite identification, pathway analysis, applications of metabolomics and its future. *Journal of mass spectrometry*, 51(8):535-548.

Barré-Sinoussi, F., Chermann, J.-C., Rey, F., Nugeyre, M.T., Chamaret, S., Gruest, J., Dauguet, C., Axler-Blin, C., Vézinet-Brun, F. & Rouzioux, C. 1983. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science*, 220(4599):868-871.

Barreto-de-Souza, V., Arakelyan, A., Margolis, L. & Vanpouille, C. 2014. HIV-1 vaginal transmission: cell-free or cell-associated virus? *American journal of reproductive immunology (New York, N.Y. : 1989)*, 71(6):589-599.

Benjamini, Y. & Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289-300.

Blankson, J.N., Persaud, D. & Siliciano, R.F. 2002. The challenge of viral reservoirs in HIV-1 infection. *Annual review of medicine*, 53(1):557-593.

Borghans, J.A.M., Mølgaard, A., de Boer, R.J. & Keşmir, C. 2007. HLA Alleles Associated with Slow Progression to AIDS Truly Prefer to Present HIV-1 p24. *PLOS ONE*, 2(9):e920.

Broquist, H.P. 1991. Lysine-pipecolic acid metabolic relationships in microbes and mammals. *Annual Review of Nutrition*, 11:435-448.

Brumme, Z., Wang, B., Nair, K., Brumme, C., de Pierres, C., Reddy, S., Julg, B., Moodley, E., Thobakgale, C. & Lu, Z. 2009. Impact of select immunologic and virologic biomarkers on CD4 cell count decrease in patients with chronic HIV-1 subtype C infection: results from Sinikithemba Cohort, Durban, South Africa. *Clinical infectious diseases*, 49(6):956-964.

Carey, B.S., Poulton, K.V. & Poles, A. 2019. Factors affecting HLA expression: A review. *International journal of immunogenetics*, 46:307– 320.

Carlson, J.M., Listgarten, J., Pfeifer, N., Tan, V., Kadie, C., Walker, B.D., Ndung'u, T., Shapiro, R., Frater, J. & Brumme, Z.L. 2012. Widespread impact of HLA restriction on immune control and escape pathways of HIV-1. *Journal of virology*, 86(9):5230-5243.

Carrington, M. & Walker, B.D. 2012. Immunogenetics of spontaneous control of HIV. *Annual Review of Medicine*, 63:131-145.

Cassol, E., Misra, V., Dutta, A., Morgello, S. & Gabuzda, D. 2014. Cerebrospinal fluid metabolomics reveals altered waste clearance and accelerated aging in HIV patients with neurocognitive impairment. *Aids*, 28(11):1579-1591.

Cassol, E., Misra, V., Holman, A., Kamat, A., Morgello, S. & Gabuzda, D. 2013. Plasma metabolomics identifies lipid abnormalities linked to markers of inflammation, microbial translocation, and hepatic function in HIV patients receiving protease inhibitors. *BMC Infectious Diseases*, 13:203.

Castellano, P., Prevedel, L., Valdebenito, S. and Eugenin, E.A., 2019. HIV infection and latency induce a unique metabolic signature in human macrophages. Scientific reports, 9(1), pp.1-14.

Chatterjee, K. 2010. Host genetic factors in susceptibility to HIV-1 infection and progression to AIDS. *Journal of Genetics*, 89(1):109-116.

Chetwynd, A.J., Samarawickrama, A., Vera, J.H., Bremner, S.A., Abdul-Sada, A., Gilleece, Y., Holt, S.G. & Hill, E.M. 2017. Nanoflow-Nanospray Mass Spectrometry Metabolomics Reveals Disruption of the Urinary Metabolite Profiles of HIV-Positive Patients on Combination Antiretroviral Therapy. *Journal of acquired immune deficiency syndromes*, 74(2):e45-e53.

Chong, J., Soufan, O., Li, C., Caraus, I., Li, S., Bourque, G., Wishart, D.S. & Xia, J. 2018. MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. *Nucleic acids research*, 46(W1):W486-W494.

Chun, T.-W. & Fauci, A.S. 1999. Latent reservoirs of HIV: Obstacles to the eradication of virus. *Proceedings of the National Academy of Sciences*, 96(20):10958-10961.

Cohen, J. 1988. The effect size index: d. *Statistical power analysis for the behavioral sciences*, 2:284-288.

Cohen, M.S., Shaw, G.M., McMichael, A.J. & Haynes, B.F. 2011. Acute HIV-1 infection. *New England Journal of Medicine*, 364(20):1943-1954.

Conway, M.E. & Hutson, S.M. 2016. *(In* Schousboe, A. & Sonnewald, U., *eds*. The Glutamate/GABA-Glutamine Cycle: Amino Acid Neurotransmitter Homeostasis. Cham: Springer International Publishing. p. 99-132).

Costello, C., Tang, J., Rivers, C., Karita, E., Meizen-Derr, J., Allen, S. & Kaslow, R.A. 1999. HLA-B* 5703 independently associated with slower HIV-1 disease progression in Rwandan women. *Aids*, 13(14):1990.

Dagenais-Lussier, X., Mouna, A., Routy, J.P., Tremblay, C., Sekaly, R.P., El-Far, M. & Grevenynghe, J. 2015. Current topics in HIV-1 pathogenesis: The emergence of deregulated immuno-metabolism in HIV-infected subjects. *Cytokine Growth Factor Reviews*, 26(6):603-613.

Department of Health. 2016. Implementation of the universal test and treat strategy for hiv positive patients and differentiated care for stable patients [1 September 2016].

Dettmer, K., Aronov, P.A. & Hammock, B.D. 2007. Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews*, 26(1):51-78.

Dickens, A.M., Anthony, D.C., Deutsch, R., Mielke, M.M., Claridge, T.D., Grant, I., Franklin, D., Rosario, D., Marcotte, T. & Letendre, S. 2015. Cerebrospinal fluid metabolomics implicate bioenergetic adaptation as a neural mechanism regulating shifts in cognitive states of HIV-infected patients. *AIDS (London, England)*, 29(5):559.

Dunn, W.B., Bailey, N.J. & Johnson, H.E. 2005. Measuring the metabolome: current analytical technologies. *Analyst*, 130(5):606-625.

Dunn, W.B. & Ellis, D.I. 2005. Metabolomics: Current analytical platforms and methodologies. *TrAC Trends in Analytical Chemistry*, 24(4):285-294.

Epstein, A.A., Narayanasamy, P., Dash, P.K., High, R., Bathena, S.P.R., Gorantla, S., Poluektova, L.Y., Alnouti, Y., Gendelman, H.E. & Boska, M.D. 2013. Combinatorial assessments of brain tissue metabolomics and histopathology in rodent models of human immunodeficiency virus infection. *Journal of Neuroimmune Pharmacology*, 8(5):1224-1238.

F. Hair Jr, J., Sarstedt, M., Hopkins, L. & G. Kuppelwieser, V. 2014. Partial least squares structural equation modeling (PLS-SEM) An emerging tool in business research. *European Business Review*, 26(2):106-121.

Fearon, K., Strasser, F., Anker, S.D., Bosaeus, I., Bruera, E., Fainsinger, R.L., Jatoi, A., Loprinzi, C., MacDonald, N., Mantovani, G., Davis, M., Muscaritoli, M., Ottery, F., Radbruch, L., Ravasco, P., Walsh, D., Wilcock, A., Kaasa, S. & Baracos, V.E. 2011. Definition and classification of cancer cachexia: an international consensus. *The Lancet Oncology*, 12(5):489-495.

Fellay, J., Ge, D., Shianna, K.V., Colombo, S., Ledergerber, B., Cirulli, E.T., Urban, T.J., Zhang, K., Gumbs, C.E., Smith, J.P., Castagna, A., Cozzi-Lepri, A., De Luca, A., Easterbrook, P., Günthard, H.F., Mallal, S., Mussini, C., Dalmau, J., Martinez-Picado, J., Miro, J.M., Obel, N., Wolinsky, S.M., Martinson, J.J., Detels, R., Margolick, J.B., Jacobson, L.P., Descombes, P., Antonarakis, S.E., Beckmann, J.S., O'Brien, S.J., Letvin, N.L., McMichael, A.J., Haynes, B.F., Carrington, M., Feng, S., Telenti, A., Goldstein, D.B. & Immunology, N.C.f.H.A.V. 2009. Common Genetic Variation and the Control of HIV-1 in Humans. *PLOS Genetics*, 5(12):e1000791.

Fernandes-Cardoso, J., Suffert, T.A., Correa Mda, G., Jobim, L.F., Jobim, M., Salim, P.H., Arruda, M.B., Boullosa, L.T., Tanuri, A., Porto, L.C. & Ferreira, O.C., Jr. 2016. Association between KIR genotypes and HLA-B alleles on viral load in Southern Brazilian individuals infected by HIV-1 subtypes B and C. *Human Immunology*, 77(10):854-860.

Fernstrom, J.D. & Fernstrom, M.H. 2007. Tyrosine, phenylalanine, and catecholamine synthesis and function in the brain. *Journal of Nutrition*, 137(6):1539-1547.

Fiehn, O. 2002. Metabolomics—the link between genotypes and phenotypes, In: Functional genomics. Springer. p. 155-171).

Fiehn, O. 2016. Metabolomics by gas chromatography–mass spectrometry: combined targeted and untargeted profiling. *Current protocols in molecular biology*:30.34. 31-30.34. 32.

Field, A. 2000. Discovering statistics using SPSS:(and sex, drugs and rock'n'roll). Pages: 497: Sage.

Fitzpatrick, M. & Young, S.P. 2013. Metabolomics – A novel window into inflammatory disease. *Swiss medical weekly*, 143:w13743.

Flores-Villanueva, P.O., Yunis, E.J., Delgado, J.C., Vittinghoff, E., Buchbinder, S., Leung, J.Y., Uglialoro, A.M., Clavijo, O.P., Rosenberg, E.S., Kalams, S.A., Braun, J.D., Boswell, S.L., Walker, B.D. & Goldfeld, A.E. 2001. Control of HIV-1 viremia and protection from AIDS are associated with HLA-Bw4 homozygosity. *Proceedings of the National Academy of Sciences of the United States of America*, 98(9):5140-5145.

Fraser, C., Hollingsworth, T.D., Chapman, R., de Wolf, F. & Hanage, W.P. 2007. Variation in HIV-1 set-point viral load: epidemiological analysis and an evolutionary hypothesis. *Proceedings of the National Academy of Sciences*, 104(44):17441-17446.

Fuchs, D., Forsman, A., Hagberg, L., Larsson, M., Norkrans, G., Reibnegger, G., Werner, E. & Wachter, H. 1990. Immune activation and decreased tryptophan in patients with HIV-1 infection. *Journal of interferon research*, 10(6):599-603.

Gao, F., Bailes, E., Robertson, D.L., Chen, Y., Rodenburg, C.M., Michael, S.F., Cummins, L.B., Arthur, L.O., Peeters, M. & Shaw, G.M. 1999. Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. *Nature*, 397(6718):436.

Gasper, D.J., Tejera, M.M. & Suresh, M. 2014. CD4 T-cell memory generation and maintenance. *Critical reviews in immunology*, 34(2):121-146.

Gentle, N.L., Loubser, S., Paximadis, M., Puren, A. & Tiemessen, C.T. 2017. Killer-cell immunoglobulin-like receptor (KIR) and human leukocyte antigen (HLA) class I genetic diversity in four South African populations. *Human immunology*, 78(7-8):503-509.

Ghannoum, M.A., Mukherjee, P.K., Jurevic, R.J., Retuerto, M., Brown, R.E., Sikaroodi, M., Webster-Cyriaque, J. & Gillevet, P.M. 2013. Metabolomics reveals differential levels of oral metabolites in HIV-infected patients: toward novel diagnostic targets. *Omics: a journal of integrative biology*, 17(1):5-15.

Goovaerts, O. 2015. Pathogenesis of tuberculosis-associated immune reconstitution inflammatory syndrome (TB-IRIS) - The calm before the cytokine storm. Antwerp: Universiteit Antwerpen.

Gostner, J.M., Becker, K., Kurz, K. & Fuchs, D. 2015a. Disturbed Amino Acid Metabolism in HIV: Association with Neuropsychiatric Symptoms. *Frontiers in Psychiatry*, 6(97).

Gottlieb, M.S., Schroff, R., Schanker, H.M., Weisman, J.D., Fan, P.T., Wolf, R.A. & Saxon, A. 1981. Pneumocystis carinii pneumonia and mucosal candidiasis in previously healthy homosexual men: evidence of a new acquired cellular immunodeficiency. *New England Journal of Medicine*, 305(24):1425-1431.

Grabar, S., Selinger-Leneman, H., Abgrall, S., Pialoux, G., Weiss, L. & Costagliola, D. 2009. Prevalence and comparative characteristics of long-term nonprogressors and HIV controller patients in the French Hospital Database on HIV. *Aids*, 23(9):1163-1169.

Grabar, S., Selinger-Leneman, H., Abgrall, S., Pialoux, G., Weiss, L. & Costagliola, D. 2017. Loss of long-term non-progressor and HIV controller status over time in the French Hospital Database on HIV-ANRS CO4. *PloS one*, 12(10):e0184441.

Graber, A.L. 2001. Syndrome of lipodystrophy, hyperlipidemia, insulin resistance, and diabetes in treated patients with human immunodeficiency virus infection. *Endocrine Practice*, 7(6):430-437.

Gu, H., Pan, Z., Xi, B., Hainline, B.E., Shanaiah, N., Asiago, V., Gowda, G.N. & Raftery, D. 2009. 1H NMR metabolomics study of age profiling in children. *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In vivo*, 22(8):826-833.

Gupta, V. & Gupta, S. 2004. Laboratory markers associated with progression of HIV infection. *Indian journal of medical microbiology*, 22(1):7.

Hahn, J.A., Cheng, D.M., Emenyonu, N.I., Lloyd-Travaglini, C., Fatch, R., Shade, S.B., Ngabirano, C., Adong, J., Bryant, K. & Muyindike, W.R. 2018. Alcohol Use and HIV Disease Progression in an Antiretroviral Naive Cohort. *Journal of acquired immune deficiency syndromes (1999)*, 77(5):492-501.

Hazenberg, M.D., Otto, S.A., van Benthem, B.H., Roos, M.T., Coutinho, R.A., Lange, J.M., Hamann, D., Prins, M. & Miedema, F. 2003. Persistent immune activation in HIV-1 infection is associated with progression to AIDS. *AIDS*, 17(13):1881-1888.

Hegedus, A., Williamson, M.K. & Huthoff, H. 2014. HIV-1 pathogenicity and virion production are dependent on the metabolic phenotype of activated CD4+ T cells. *Retrovirology*, 11(1):98.

Hemelaar, J. 2012. The origin and diversity of the HIV-1 pandemic. *Trends in molecular medicine*, 18(3):182-192.

Herman, M.A., She, P., Peroni, O.D., Lynch, C.J. & Kahn, B.B. 2010. Adipose tissue branched chain amino acid (BCAA) metabolism modulates circulating BCAA levels. *Journal of Biological Chemistry*, 285(15):11348-11356.

Herrera, S., Martínez-Sanz, J. & Serrano-Villar, S. 2019. HIV, Cancer, and the Microbiota: Common Pathways Influencing Different Diseases. *Frontiers in immunology*, 10:1466-1466.

Hewer, R., Vorster, J., Steffens, F. & Meyer, D. 2006. Applying biofluid 1H NMR-based metabonomic techniques to distinguish between HIV-1 positive/AIDS patients on antiretroviral treatment and HIV-1 negative individuals. *Journal of pharmaceutical and biomedical analysis*, 41(4):1442-1446.

HIV/AIDS, J.U.N.P.o. 2017. Ending AIDS: Progress towards the 90-90-90 targets. *Global AIDS update*.

Holecek, M. 2002. Relation between glutamine, branched-chain amino acids, and protein metabolism. *Nutrition*, 18(2):130-133.

Holeček, M. 2018. Branched-chain amino acids in health and disease: metabolism, alterations in blood plasma, and as supplements. *Nutrition & metabolism*, 15:33-33.

Hollenbaugh, J.A., Munger, J. & Kim, B. 2011. Metabolite profiles of human immunodeficiency virus infected CD4+ T cells and macrophages using LC–MS/MS analysis. *Virology*, 415(2):153-159.

Hommes, M.J., Romijn, J.A., Endert, E. & Sauerwein, H.P. 1991. Resting energy expenditure and substrate oxidation in human immunodeficiency virus (HIV)-infected asymptomatic men: HIV affects host metabolism in the early asymptomatic stage. *The American Journal of Clinical Nutrition*, 54(2):311-315.

Hsu, J.W., Pencharz, P.B., Macallan, D. & Tomkins, A. 2005. Macronutrients and HIV/AIDS: a review of current evidence. *Durban, South Africa: World Health Organization*.

Hu, W.-S. & Hughes, S.H. 2012. HIV-1 reverse transcription. *Cold Spring Harbor perspectives in medicine*, 2(10):a006882.

Huang, Y., Paxton, W.A., Wolinsky, S.M., Neumann, A.U., Zhang, L., He, T., Kang, S., Ceradini, D., Jin, Z. & Yazdanbakhsh, K. 1996. The role of a mutant CCR5 allele in HIV–1 transmission and disease progression. *Nature medicine*, 2(11):1240.

Hughes, M.D., Stein, D.S., Gundacker, H.M., Valentine, F.T., Phair, J.P. & Volberding, P.A. 1994. Within-subject variation in CD4 lymphocyte count in asymptomatic human immunodeficiency virus infection: implications for patient monitoring. *Journal of Infectious Diseases*, 169(1):28-36.

Hurley, R. 2018. HIV/AIDS: complacency risks reversing progress on ending epidemic, conference hears. *BMJ*, 362:k3241.

Ipp, H., Zemlin, A.E., Erasmus, R.T. & Glashoff, R.H. 2014. Role of inflammation in HIV-1 disease progression and prognosis. *Critical Reviews in Clinical Laboratory Sciences*, 51(2):98-111.

Ivanisevic, J., Elias, D., Deguchi, H., Averell, P.M., Kurczy, M., Johnson, C.H., Tautenhahn, R., Zhu, Z., Watrous, J., Jain, M., Griffin, J., Patti, G.J. & Siuzdak, G. 2015. Arteriovenous Blood Metabolomics: A Readout of Intra-Tissue Metabostasis. *Scientific Reports*, 5:12757.

Jenabian, M.-A., Patel, M., Kema, I., Kanagaratham, C., Radzioch, D., Thébault, P., Lapointe, R., Tremblay, C., Gilmore, N. & Ancuta, P. 2013. Distinct tryptophan catabolism and Th17/Treg balance in HIV progressors and elite controllers. *PloS one*, 8(10):e78146.

Jiang, Y., Chen, O., Cui, C., Zhao, B., Han, X., Zhang, Z., Liu, J., Xu, J., Hu, Q., Liao, C. & Shang, H. 2013. KIR3DS1/L1 and HLA-Bw4-80I are associated with HIV disease progression among HIV typical progressors and long-term nonprogressors. *BMC Infectious Diseases*, 13:405.

Johnson, C.H. & Gonzalez, F.J. 2012. Challenges and opportunities of metabolomics. *Journal of cellular physiology*, 227(8):2975-2981.

Joint United Nations Programme on HIV/AIDS. 2014. 90-90-90: an ambitious treatment target to help end the AIDS epidemic. *Geneva: Unaids*.

Jornayvaz, F.R. & Shulman, G.I. 2010. Regulation of mitochondrial biogenesis. *Essays in biochemistry*, 47:69-84.

Justice, A., McGinnis, K., Skanderson, M., Chang, C., Gibert, C., Goetz, M., Rimland, D., Rodriguez-Barradas, M., Oursler, K. & Brown, S. 2010. Towards a combined prognostic index for survival in HIV infection: the role of ‘non-HIV’biomarkers. *HIV medicine*, 11(2):143-151.

Kamleh, M.A., Hobani, Y., Dow, J.A., Zheng, L. & Watson, D.G. 2009. Towards a platform for the metabonomic profiling of different strains of Drosophila melanogaster using liquid chromatography–Fourier transform mass spectrometry. *The FEBS journal*, 276(22):6798-6809.

Klein, F., Halper-Stromberg, A., Horwitz, J.A., Gruell, H., Scheid, J.F., Bournazos, S., Mouquet, H., Spatz, L.A., Diskin, R. & Abadir, A. 2012. HIV therapy by a combination of broadly neutralizing antibodies in humanized mice. *Nature*, 492(7427):118.

Kløverpris, H.N., Harndahl, M., Leslie, A.J., Carlson, J.M., Ismail, N., van der Stok, M., Huang, K.-H.G., Chen, F., Riddell, L. & Steyn, D. 2012. HIV control through a single nucleotide on the HLA-B locus. *Journal of virology*, 86(21):11493-11500.

Knapp, D.R. 1979. Handbook of analytical derivatization reactions: John Wiley & Sons.

Koethe, J.R., Grome, H., Jenkins, C.A., Kalams, S.A. & Sterling, T.R. 2016. The metabolic and cardiovascular consequences of obesity in persons with HIV on long-term antiretroviral therapy. *AIDS (London, England)*, 30(1):83.

Korenromp, E.L., Williams, B.G., Schmid, G.P. & Dye, C. 2009. Clinical prognostic value of RNA viral load and CD4 cell counts during untreated HIV-1 infection—a quantitative review. *PloS one*, 4(6):e5950.

Korzeniewski, B. 2001. Theoretical studies on the regulation of oxidative phosphorylation in intact tissues. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1504(1):31-45.

Kulkarni, M.M., Ratcliff, A.N., Bhat, M., Alwarawrah, Y., Hughes, P., Arcos, J., Loiselle, D., Torrelles, J.B., Funderburg, N.T., Haystead, T.A. & Kwiek, J.J. 2017. Cellular fatty acid synthase is required for late stages of HIV-1 replication. *Retrovirology*, 14(1):45.

Kuller, L.H., Tracy, R., Belloso, W., De Wit, S., Drummond, F., Lane, H.C., Ledergerber, B., Lundgren, J., Neuhaus, J. & Nixon, D. 2008. Inflammatory and coagulation biomarkers and mortality in patients with HIV infection. *PLoS medicine*, 5(10):e203.

Kumar, P. 2013. Long term non-progressor (LTNP) HIV infection. *Indian Journal of Medical Research*, 138(3):291-293.

Lake, J.E. & Currier, J.S. 2013. Metabolic disease in HIV infection. *The Lancet infectious diseases*, 13(11):964-975.

Lange, J. & Ananworanich, J. 2014. The discovery and development of antiretroviral agents. *Antiviral Therapy*, 19(Suppl 3):5-14.

Lange, N., Carlin, B.P. & Gelfand, A.E. 1992. Hierarchical Bayes models for the progression of HIV infection using longitudinal CD4 T-cell numbers. *Journal of the American Statistical Association*, 87(419):615-626.

Langford, S.E., Ananworanich, J. & Cooper, D.A. 2007. Predictors of disease progression in HIV infection: a review. *AIDS research and therapy*, 4(1):11.

Langkilde, A., Petersen, J., Henriksen, J.H., Jensen, F.K., Gerstoft, J., Eugen-Olsen, J. & Andersen, O. 2015. Leptin, IL-6, and suPAR reflect distinct inflammatory changes associated with adiposity, lipodystrophy and low muscle mass in HIV-infected patients and controls. *Immune system and aging*, 12:9.

Leserman, J. 2000. The effects of depression, stressful life events, social support, and coping on the progression of HIV infection. *Current Psychiatry Reports*, 2(6):495-502.

Leserman, J., Jackson, E.D., Petitto, J.M., Golden, R.N., Silva, S.G., Perkins, D.O., Cai, J., Folds, J.D. & Evans, D.L. 1999. Progression to AIDS: The Effects of Stress, Depressive Symptoms, and Social Support. *Psychosomatic Medicine*, 61(3):397-406.

Leserman, J., Petitto, J., Gu, H., Gaynes, B., Barroso, J., Golden, R., Perkins, D., Folds, J. & Evans, D. 2002. Progression to AIDS, a clinical AIDS condition and mortality: psychosocial and physiological predictors. *Psychological medicine*, 32(6):1059-1073.

Leslie, A., Matthews, P.C., Listgarten, J., Carlson, J.M., Kadie, C., Ndung'u, T., Brander, C., Coovadia, H., Walker, B.D. & Heckerman, D. 2010. Additive contribution of HLA class I alleles in the immune control of HIV-1 infection. *Journal of virology*, 84(19):9879-9888.

Limou, S. & Zagury, J.F. 2013. Immunogenetics: Genome-Wide Association of Non-Progressive HIV and Viral Load Control: HLA Genes and Beyond. *Frontiers in Immunology*, 4:118.

Liovat, A.-S., Rey-Cuillé, M.-A., Lécuroux, C., Jacquelin, B., Girault, I., Petitjean, G., Zitoun, Y., Venet, A., Barré-Sinoussi, F. & Lebon, P. 2012. Acute plasma biomarkers of T cell activation set-point levels and of disease progression in HIV-1 infection. *PloS one*, 7(10):e46143.

Liu, Z. & Phillips, J.B. 1991. Comprehensive two-dimensional gas chromatography using an on-column thermal modulator interface. *Journal of Chromatographic Science*, 29(6):227-231.

Loubser, S. 2015. The multiple roles of hla in hiv immunity and treatment*.* Johannesburg: University of Witwatersrand.

Loubser, S., Paximadis, M. & Tiemessen, C.T. 2017. Human leukocyte antigen class I (A, B and C) allele and haplotype variation in a South African Mixed ancestry population. *Human immunology*, 78(5-6):399-400.

Lu, W., Chen, S., Lai, C., Lai, M., Fang, H., Dao, H., Kang, J., Fan, J., Guo, W., Fu, L. & Andrieu, J.M. 2016. Suppression of HIV Replication by CD8(+) Regulatory T-Cells in Elite Controllers. *Frontiers in Immunology*, 7:134.

Ma, N. & Ma, X. 2019. Dietary Amino Acids and the Gut-Microbiome-Immune Axis: Physiological Metabolism and Therapeutic Prospects. *Comprehensive Reviews in Food Science and Food Safety*, 18(1):221-242.

MacNamara, K., Leardi, R. & Hoffmann, A. 2004. Developments in 2-D gas chromatography with heartcutting. *LC GC North America*, 22(9; SUPP):82-91.

Madec, Y., Boufassa, F., Avettand-Fenoel, V., Hendou, S., Melard, A., Boucherit, S., Surzyn, J., Meyer, L., Rouzioux, C. & Group, A.S.H.S. 2009. Early control of HIV-1 infection in long-term nonprogressors followed since diagnosis in the ANRS SEROCO/HEMOCO cohort. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 50(1):19-26.

Madhu, B., Narita, M., Jauhiainen, A., Menon, S., Stubbs, M., Tavaré, S., Narita, M. & Griffiths, J.R. 2015. Metabolomic changes during cellular transformation monitored by metabolite–metabolite correlation analysis and correlated with gene expression. *Metabolomics*, 11(6):1848-1863.

Maher, A.D., Cysique, L.A., Brew, B.J. & Rae, C.D. 2011. Statistical integration of 1H NMR and MRS data from different biofluids and tissues enhances recovery of biological information from individuals with HIV-1 infection. *Journal of proteome research*, 10(4):1737-1745.

Mandalia, S., Westrop, S., Beck, E., Nelson, M., Gazzard, B. & Imami, N. 2011. Frequency and characteristics of long-term non-progressors and HIV controllers in the Chelsea and Westminster HIV cohort. *HIV Medicine*, 12:10-11.

Marriott, P. & Nolvachai, Y. 2018. Resolution Revolution: Can GC× GC be considered a super-resolution technique? *The Analytical Scientist*, 67.

Masson, J.J.R., Murphy, A.J., Lee, M.K.S., Ostrowski, M., Crowe, S.M. & Palmer, C.S. 2017. Assessment of metabolic and mitochondrial dynamics in CD4+ and CD8+ T cells in virologically suppressed HIV-positive individuals on combination antiretroviral therapy. *PLOS ONE*, 12(8):e0183931.

Matthews, P.C., Adland, E., Listgarten, J., Leslie, A., Mkhwanazi, N., Carlson, J.M., Harndahl, M., Stryhn, A., Payne, R.P. & Ogwu, A. 2011. HLA-A* 7401–mediated control of HIV viremia is independent of its linkage disequilibrium with HLA-B* 5703. *The Journal of Immunology*, 186(10):5675-5686.

McKnight, T.R., Yoshihara, H.A., Sitole, L.J., Martin, J.N., Steffens, F. & Meyer, D. 2014. A combined chemometric and quantitative NMR analysis of HIV/AIDS serum discloses metabolic alterations associated with disease status. *Molecular BioSystems*, 10(11):2889-2897.

McMichael, A.J., Borrow, P., Tomaras, G.D., Goonetilleke, N. & Haynes, B.F. 2009. The immune response during acute HIV-1 infection: clues for vaccine development. *Nature Reviews Immunology*, 10:11.

Mega, E.R. 2019. 'Mosaic' HIV vaccine to be tested in thousands of people across the world. *Nature*, 572(7768):165-166.

Mellors, J.W., Margolick, J.B., Phair, J.P., Rinaldo, C.R., Detels, R., Jacobson, L.P. & Muñoz, A. 2007. Prognostic value of HIV-1 RNA, CD4 cell count, and CD4 Cell count slope for progression to AIDS and death in untreated HIV-1 infection. *Jama*, 297(21):2346-2350.

Miura, T., Brockman, M.A., Brumme, C.J., Brumme, Z.L., Carlson, J.M., Pereyra, F., Trocha, A., Addo, M.M., Block, B.L. & Rothchild, A.C. 2008. Genetic characterization of human immunodeficiency virus type 1 in elite controllers: lack of gross genetic defects or common amino acid changes. *Journal of virology*, 82(17):8422-8430.

Mlawanda, G., Rheeder, P. & Miot, J. 2012. Inter-and intra-laboratory variability of CD4 cell counts in Swaziland. *Southern African Journal of HIV Medicine*, 13(2).

Mlisana, K., Werner, L., Garrett, N.J., McKinnon, L.R., van Loggerenberg, F., Passmore, J.A., Gray, C.M., Morris, L., Williamson, C., Abdool Karim, S.S. & Centre for the, A.P.o.R.i.S.A.S.T. 2014. Rapid disease progression in HIV-1 subtype C-infected South African women. *Clinical Infectious Diseases*, 59(9):1322-1331.

Monirujjaman, M. & Ferdouse, A. 2014. Metabolic and physiological roles of branched-chain amino acids. *Advances in Molecular Biology*, 2014.

Moutloatse, G.P., Bunders, M.J., van Reenen, M., Mason, S., Kuijpers, T.W., Engelke, U.F., Wevers, R.A. & Reinecke, C.J. 2016. Metabolic risks at birth of neonates exposed in utero to

HIV-antiretroviral therapy relative to unexposed neonates: an NMR metabolomics study of cord blood. *Metabolomics*, 12(11):175.

Mulligan, K., Grunfeld, C., Tai, V.W., Algren, H., Pang, M., Chernoff, D.N., Lo, J.C. & Schambelan, M. 2000. Hyperlipidemia and insulin resistance are induced by protease inhibitors independent of changes in body composition in patients with HIV infection. *Journal of acquired immune deficiency syndromes (1999)*, 23(1):35-43.

Munshi, S.U., Rewari, B.B., Bhavesh, N.S. & Jameel, S. 2013. Nuclear magnetic resonance based profiling of biofluids reveals metabolic dysregulation in HIV-infected persons and those on anti-retroviral therapy. *PloS one*, 8(5):e64298.

N., K.H., Emily, A., Madoka, K., Anette, S., Mikkel, H., C., M.P., Roger, S., D., W.B., Thumbi, N.u., Christian, B., Masafumi, T., Søren, B. & Philip, G. 2014. HIV Subtype Influences HLA-B*07:02-Associated HIV Disease Outcome. *AIDS Research and Human Retroviruses*, 30(5):468-475.

Neefjes, J., Jongsma, M.L.M., Paul, P. & Bakke, O. 2011. Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nature Reviews Immunology*, 11:823.

Nixon, D.E. & Landay, A.L. 2010. Biomarkers of immune dysfunction in HIV. *Current Opinion in HIV and AIDS*, 5(6):498.

Nonodi, T.P. & Meyer, D. 2014. NMR Metabonomics in an In Vitro Model of HIV-1 Latency. *AIDS research and human retroviruses*, 30(S1):A144-A144.

Norton, L.E. & Layman, D.K. 2006. Leucine regulates translation initiation of protein synthesis in skeletal muscle after exercise. *The Journal of nutrition*, 136(2):533S-537S.

O'Connell, K.A., Rabi, S.A., Siliciano, R.F. & Blankson, J.N. 2011. CD4+ T cells from elite suppressors are more susceptible to HIV-1 but produce fewer virions than cells from chronic progressors. *Proceedings of the National Academy of Sciences*, 108(37):E689-E698.

Olivier, I. & Loots, D.T. 2012. A comparison of two extraction methods for differentiating and characterising various Mycobacterium species and Pseudomonas aeruginosa using GC-MS metabolomics. *African Journal of Microbiology Research*, 6(13):3159-3172.

Olson, A.D., Meyer, L., Prins, M., Thiebaut, R., Gurdasani, D., Guiguet, M., Chaix, M.L., Amornkul, P., Babiker, A., Sandhu, M.S., Porter, K. & EuroCoord, C.C.i. 2014. An evaluation of HIV elite controller definitions within a large seroconverter cohort collaboration. *PLoS One*, 9(1):e86719.

Olvera, A., Pérez-Alvarez, S., Ganoza, C., Lama, J., Bernard, N., Sanchez, J. & Brander, C. 2014. Combined effect of HLA-C* 04: 01 and KIR2DS4 on increased HIV viral loads. *AIDS research and human retroviruses*, 30(S1):A39-A40.

Paiardini, M., Frank, I., Pandrea, I., Apetrei, C. & Silvestri, G. 2008. Mucosal immune dysfunction in AIDS pathogenesis. *AIDS reviews*, 10(1):36-46.

Palmer, C.S., Cherry, C.L., Sada-Ovalle, I., Singh, A. & Crowe, S.M. 2016. Glucose Metabolism in T Cells and Monocytes: New Perspectives in HIV Pathogenesis. *EBioMedicine*, 6:31-41.

Palmer, C.S., Ostrowski, M., Gouillou, M., Tsai, L., Yu, D., Zhou, J., Henstridge, D.C., Maisa, A., Hearps, A.C., Lewin, S.R., Landay, A., Jaworowski, A., McCune, J.M. & Crowe, S.M. 2014. Increased glucose metabolic activity is associated with CD4(+) T-cell activation and depletion during chronic HIV infection. *AIDS (London, England)*, 28(3):297-309.

Pantaleo, G. & Fauci, A. 1996. Immunopathogenesis of HIV infection. *Annual Reviews in Microbiology*, 50(1):825-854.

Parihar, S.P., Ozturk, M., Marakalala, M.J., Loots, D.T., Hurdayal, R., Maasdorp, D.B., Van Reenen, M., Zak, D.E., Darboe, F., Penn-Nicholson, A., Hanekom, W.A., Leitges, M., Scriba, T.J., Guler, R. & Brombacher, F. 2017. Protein kinase C-delta (PKCδ), a marker of inflammation and tuberculosis disease progression in humans, is important for optimal macrophage killing effector functions and survival in mice. *Mucosal Immunology*, 11:496.

Payne, R., Muenchhoff, M., Mann, J., Roberts, H.E., Matthews, P., Adland, E., Hempenstall, A., Huang, K.-H., Brockman, M., Brumme, Z., Sinclair, M., Miura, T., Frater, J., Essex, M., Shapiro, R., Walker, B.D., Ndung'u, T., McLean, A.R., Carlson, J.M. & Goulder, P.J.R. 2014. Impact of HLA-driven HIV adaptation on virulence in populations of high HIV seroprevalence. *Proceedings of the National Academy of Sciences*, 111(50):E5393-E5400.

Peeters, M., Toure-Kane, C. & Nkengasong, J.N. 2003. Genetic diversity of HIV in Africa: impact on diagnosis, treatment, vaccine development and trials. *AIDS*, 17(18):2547-2560.

Pelley, J.W. 2012. *(In* Pelley, J.W., *ed.* Elsevier's Integrated Review Biochemistry (Second Edition). Philadelphia: W.B. Saunders. p. 109-117).

Pendyala, G. & Fox, H.S. 2010. Proteomic and metabolomic strategies to investigate HIV-associated neurocognitive disorders. *Genome medicine*, 2(3):22.

Peng, D.X. & Lai, F. 2012. Using partial least squares in operations management research: A practical guideline and summary of past research. *Journal of Operations Management*, 30(6):467-480.

Perfetto, S.P., Chattopadhyay, P.K. and Roederer, M., 2004. Seventeen-colour flow cytometry: unravelling the immune system. Nature Reviews Immunology, 4(8), pp.648-655.

Philippeos, C., Steffens, F. & Meyer, D. 2009. Comparative 1H NMR-based metabonomic analysis of HIV-1 sera. *Journal of biomolecular NMR*, 44(3):127-137.

Post, F., Wood, R. & Maartens, G. 1996. CD4 and total lymphocyte counts as predictors of HIV disease progression. *QJM: An International Journal of Medicine*, 89(7):505-508.

Raboud, J., Haley, L., Montaner, J., Murphy, C., Januszewska, M. & Schechter, M. 1995. Quantification of the variation due to laboratory and physiologic sources in CD4 lymphocyte counts of clinically stable HIV-infected individuals. *Journal of acquired immune deficiency syndromes and human retrovirology: official publication of the International Retrovirology Association*, 10:S67-73.

Rappocciolo, G., Jais, M., Piazza, P., Reinhart, T.A., Berendam, S.J., Garcia-Exposito, L., Gupta, P. & Rinaldo, C.R. 2014. Alterations in cholesterol metabolism restrict HIV-1 trans infection in nonprogressors. *MBio*, 5(3):e01031-01013.

Roberts, L.D., Souza, A.L., Gerszten, R.E. & Clish, C.B. 2012. Targeted metabolomics. *Current protocols in molecular biology*, 98(1):30.32. 31-30.32. 24.

Robinson, J., Halliwell, J.A., Hayhurst, J.D., Flicek, P., Parham, P. & Marsh, S.G. 2014. The IPD and IMGT/HLA database: allele variant databases. *Nucleic acids research*, 43(D1):D423-D431.

Rodríguez-Gallego, E., Gómez, J., Domingo, P., Ferrando-Martínez, S., Peraire, J., Viladés, C., Veloso, S., López-Dupla, M., Beltrán-Debón, R. & Alba, V. 2018a. Circulating metabolomic profile can predict dyslipidemia in HIV patients undergoing antiretroviral therapy. *Atherosclerosis*, 273:28-36.

Rodríguez-Gallego, E., Gómez, J., Pacheco, Y.M., Peraire, J., Viladés, C., Beltrán-Debón, R., Mallol, R., López-Dupla, M., Veloso, S. & Alba, V. 2018c. A baseline metabolomic signature is associated with immunological CD4+ T-cell recovery after 36 months of antiretroviral therapy in HIV-infected patients. *AIDS (London, England)*, 32(5):565.

Rousseau, C.M., Daniels, M.G., Carlson, J.M., Kadie, C., Crawford, H., Prendergast, A., Matthews, P., Payne, R., Rolland, M. & Raugi, D.N. 2008. HLA class I-driven evolution of human immunodeficiency virus type 1 subtype c proteome: immune escape and viral load. *Journal of virology*, 82(13):6434-6446.

Saag, M.S., Holodniy, M., Kuritzkes, D., O'Brien, W., Coombs, R., Poscher, M., Jacobsen, D., Shaw, G., Richman, D. & Volberding, P. 1996. HIV viral load markers in clinical practice. *Nature medicine*, 2(6):625.

Sakuma, R. & Takeuchi, H. 2012. SIV replication in human cells. *Frontiers in microbiology*, 3:162-162.

Salas-Salvado, J. & Garcia-Lorda, P. 2001. The metabolic puzzle during the evolution of HIV infection. *Clinical Nutrition*, 20(5):379-391.

Sauerwein, H.P., Poll, T. & Romijn, J.A. 2011. Cytokines: role in human metabolism. Berlin: Springer

Scarpelini, B., Zanoni, M., Sucupira, M.C.A., Truong, H.-H.M., Janini, L.M.R., Segurado, I.D.C. & Diaz, R.S. 2016. Plasma Metabolomics Biosignature According to HIV Stage of Infection, Pace of Disease Progression, Viremia Level and Immunological Response to Treatment. *PLOS ONE*, 11(12):e0161920.

Schoeman, J. & Loots, D. 2011. Improved disease characterisation and diagnostics using metabolomics: A review. *Journal of Cell and Tissue Research*, 11(1):2673.

Schoeman, J.C. & Du Preez, I. 2012. A comparison of four sputum pre-extraction preparation methods for identifying and characterising Mycobacterium tuberculosis using GCxGC-TOFMS metabolomics. *Journal of microbiological methods*, 91(2):301-311.

Schröder, A.R., Shinn, P., Chen, H., Berry, C., Ecker, J.R. & Bushman, F. 2002. HIV-1 integration in the human genome favors active genes and local hotspots. *Cell*, 110(4):521-529.

Schutte, J.E., Longhurst, J.C., Gaffney, F.A., Bastian, B.C. & Blomqvist, C.G. 1981. Total plasma creatinine: an accurate measure of total striated muscle mass. *Journal of Applied Physiology*, 51(3):762-766.

Semba, R.D., Darnton-Hill, I. & de Pee, S. 2010. Addressing tuberculosis in the context of malnutrition and HIV coinfection. *Food and nutrition bulletin*, 31(4):S345-S364.

Shi, L., Eugenin, E.A. & Subbian, S. 2016. Immunometabolism in Tuberculosis. *Frontiers in immunology*, 7:150.

Siddiqui, R.A., Sauermann, U., Altmuller, J., Fritzer, E., Nothnagel, M., Dalibor, N., Fellay, J., Kaup, F.J., Stahl-Hennig, C., Nurnberg, P., Krawczak, M. & Platzer, M. 2009. X chromosomal variation is associated with slow progression to AIDS in HIV-1-infected women. *American Journal of Human Genetics*, 85(2):228-239.

Silva, E.M., Acosta, A.X., Santos, E.J., Netto, E.M., Lemaire, D.C., Oliveira, A.S., Barbosa, C.M., Bendicho, M.T., Galvao-Castro, B. & Brites, C. 2010. HLA-Bw4-B*57 and Cw*18 alleles are associated with plasma viral load modulation in HIV-1 infected individuals in Salvador, Brazil. *Brazilian Journal of Infectious Diseases*, 14(5):468-475.

Sitole, L., Steffens, F., Krüger, T.P. & Meyer, D. 2014. Mid-ATR-FTIR spectroscopic profiling of HIV/AIDS sera for novel systems diagnostics in global health. *Omics: a journal of integrative biology*, 18(8):513-523.

Sitole, L., Steffens, F. & Meyer, D. 2015. Raman Spectroscopy-based Metabonomics of HIV-infected Sera Detects Amino Acid and Glutathione Changes. *Current Metabolomics*, 3(1):65-75.

Sitole, L.J., Williams, A.A. & Meyer, D. 2013. Metabonomic analysis of HIV-infected biofluids. *Molecular Biosystems*, 9(1):18-28.

Smith, R., Rossetto, K. & Peterson, B.L. 2008. A meta-analysis of disclosure of one's HIV-positive status, stigma and social support. *AIDS care*, 20(10):1266-1275.

Sok, D. & Burton, D.R. 2018. Recent progress in broadly neutralizing antibodies to HIV. *Nature immunology*, 19(11):1179.

Sousa, A.E., Carneiro, J., Meier-Schellersheim, M., Grossman, Z. & Victorino, R.M.M. 2002. CD4 T Cell Depletion Is Linked Directly to Immune Activation in the Pathogenesis of HIV-1 and HIV-2 but Only Indirectly to the Viral Load. *The Journal of Immunology*, 169:3400-3406.

Sullivan, G.M. & Feinn, R. 2012. Using Effect Size-or Why the P Value Is Not Enough. *Journal of graduate medical education*, 4(3):279-282.

Sundquist, W.I. & Kräusslich, H.-G. 2012. HIV-1 assembly, budding, and maturation. *Cold Spring Harbor perspectives in medicine*, 2(7):a006924.

Swanson, B., Beverly, E.S., Keithley, J.K., Fogg, L., Nerad, J., Novak, R.M. & Adeyemi, O. 2009. Lipoprotein particle profiles by nuclear magnetic resonance spectroscopy in medically underserved HIV-infected persons. *Journal of clinical lipidology*, 3(6):379-384.

Tarancon-Diez, L., Rodríguez-Gallego, E., Rull, A., Peraire, J., Viladés, C., Portilla, I., Jimenez-Leon, M.R., Alba, V., Herrero, P., Leal, M., Ruiz-Mateos, E. & Vidal, F. 2019. Immunometabolism is a key factor for the persistent spontaneous elite control of HIV-1 infection. *EBioMedicine*, 42:86-96.

Taylor, J., Fahey, J.L., Detels, R. & Giorgi, J.V. 1989. CD4 percentage, CD4 number, and CD4: CD8 ratio in HIV infection: which to choose and how to use. *Journal of acquired immune deficiency syndromes*, 2(2):114-124.

Traven, A. & Naderer, T. 2019. Central metabolic interactions of immune cells and microbes: prospects for defeating infections. *EMBO reports*:e47995.

U.S. Food and Drug Administration. 2018. Antiretroviral drugs used in the treatment of HIV infection. https://www.fda.gov/patients/hiv-treatment/antiretroviral-drugs-used-treatment-hiv-infection Date of access: 09/07/2019.

UNAIDS. 2019. HVTN 702 clinical trial of an HIV vaccine stopped. https://www.unaids.org/en/resources/presscentre/pressreleaseandstatementarchive/2020/february/20200204_vaccine Date of access: 09/04/2020.

Uribe, M.R., Adams, S., Shupert, W.L., Stroncek, D., Connors, M. & Marincola, F. 2004. Association between HLA-B57, KIR genes and HIV progression. *Human Immunology*, 65(9):S68.

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.

Van Der Kloet, F.M., Bobeldijk, I., Verheij, E.R. & Jellema, R.H. 2009. Analytical error reduction using single point calibration for accurate and precise metabolomic phenotyping. *Journal of proteome research*, 8(11):5132-5141.

van Ravenzwaay, B., Cunha, G.C.-P., Leibold, E., Looser, R., Mellert, W., Prokoudine, A., Walk, T. & Wiemer, J. 2007. The use of metabolomics for the discovery of new biomarkers of effect. *Toxicology Letters*, 172(1):21-28.

Vassimon, H.S., de Paula, F.J.A., Machado, A.A., Monteiro, J.P. & Jordão, A.A. 2012. Hypermetabolism and altered substrate oxidation in HIV-infected patients with lipodystrophy. *Nutrition*, 28(9):912-916.

Vázquez-Castellanos, J.F., Serrano-Villar, S., Jiménez-Hernández, N., del Rio, M.D.S., Gayo, S., Rojo, D., Ferrer, M., Barbas, C., Moreno, S. & Estrada, V. 2018. Interplay between gut microbiota metabolism and inflammation in HIV infection. *ISME journal*, 12(8):1964-1976.

Venner, C.M., Nankya, I., Kyeyune, F., Demers, K., Kwok, C., Chen, P.-L., Rwambuya, S., Munjoma, M., Chipato, T. & Byamugisha, J. 2016. Infecting HIV-1 subtype predicts disease progression in women of sub-Saharan Africa. *EBioMedicine*, 13:305-314.

Vesterbacka, J., Rivera, J., Noyan, K., Parera, M., Neogi, U., Calle, M., Paredes, R., Sönnerborg, A., Noguera-Julian, M. & Nowak, P. 2017. Richer gut microbiota with distinct metabolic profile in HIV infected elite controllers. *Scientific reports*, 7(1):6269.

Vujkovic-Cvijin, I., Dunham, R.M., Iwai, S., Maher, M.C., Albright, R.G., Broadhurst, M.J., Hernandez, R.D., Lederman, M.M., Huang, Y. & Somsouk, M. 2013. Dysbiosis of the gut microbiota is associated with HIV disease progression and tryptophan catabolism. *Science translational medicine*, 5(193):193ra191-193ra191.

Westerhuis, J.A., Hoefsloot, H.C., Smit, S., Vis, D.J., Smilde, A.K., van Velzen, E.J., van Duijnhoven, J.P. & van Dorsten, F.A. 2008. Assessment of PLSDA cross validation. *Metabolomics*, 4(1):81-89.

Wikoff, W.R., Pendyala, G., Siuzdak, G. & Fox, H.S. 2008. Metabolomic analysis of the cerebrospinal fluid reveals changes in phospholipase expression in the CNS of SIV-infected macaques. *The Journal of clinical investigation*, 118(7):2661-2669.

Wilen, C.B., Tilton, J.C. & Doms, R.W. 2012. *(In* Rossmann, M.G. & Rao, V.B., *eds*. Viral Molecular Machines. Boston, MA: Springer US. p. 223-242).

Williams, A., Steffens, F., Reinecke, C. & Meyer, D. 2013. The Th1/Th2/Th17 cytokine profile of HIV-infected individuals: A multivariate cytokinomics approach. *Cytokine*, 61(2):521-526.

Williams, A.A. 2012. Metabonomics profile and corresponding immune parameters of HIV infected individuals*.* Pretoria South Africa: University of Pretoria.

Wishart, D.S., Feunang, Y.D., Marcu, A., Guo, A.C., Liang, K., Vázquez-Fresno, R., Sajed, T., Johnson, D., Li, C. & Karu, N. 2017. HMDB 4.0: the human metabolome database for 2018. *Nucleic acids research*, 46(D1):D608-D617.

World Health Organisation. 2015. Consolidated guidelines on HIV testing services. https://apps.who.int/iris/bitstream/handle/10665/179870/9789241508926_eng.pdf?sequence=1 Date of access: 2019/08/12.

World Health Organization. 2013. Consolidated guidelines on the use of antiretroviral drugs for treating and preventing HIV infection: recommendations for a public health approach: World Health Organization.

World Health Organization. 2019. HIV/AIDS. https://www.who.int/news-room/fact-sheets/detail/hiv-aids Date of access: 30 Jul 2019.

Worthley, L., Philcox, J. & Hartley, T. 1984. Fasting and Non-fasting Serum Amino Acid Values in three Home Parenteral Nutrition Patients: A Comparison between Synthamin 17R and Vamin N 7% R. *Anaesthesia and intensive care*, 12(1):46-51.

Wright, J.K., Brumme, Z.L., Carlson, J.M., Heckerman, D., Kadie, C.M., Brumme, C.J., Wang, B., Losina, E., Miura, T. & Chonco, F. 2010. Gag-protease-mediated replication capacity in HIV-1 subtype C chronic infection: associations with HLA type and clinical parameters. *Journal of virology*, 84(20):10820-10831.

Xia, J. & Wishart, D.S. 2016. Using MetaboAnalyst 3.0 for comprehensive metabolomics data analysis. *Current protocols in bioinformatics*, 55(1):14.10. 11-14.10. 91.

Yu, Z., Huang, H., Reim, A., Charles, P.D., Northage, A., Jackson, D., Parry, I. & Kessler, B.M. 2017. Optimizing 2D gas chromatography mass spectrometry for robust tissue, serum and urine metabolite profiling. *Talanta*, 165:685-691.

Zangerle, R., Kurz, K., Neurauter, G., Kitchen, M., Sarcletti, M. & Fuchs, D. 2010. Increased blood phenylalanine to tyrosine ratio in HIV-1 infection and correction following effective antiretroviral therapy. *Brain, behavior, and immunity*, 24(3):403-408.

Zanoni, M., Aventurato, Í.K., Hunter, J., Sucupira, M.C.A. & Diaz, R.S. 2017. Uniquely altered transcripts are associated with immune preservation in HIV infection. *PloS one*, 12(3):e0169868.

Zarate, E., Boyle, V., Rupprecht, U., Green, S., Villas-Boas, S., Baker, P. & Pinu, F. 2017. Fully automated trimethylsilyl (TMS) derivatisation protocol for metabolite profiling by GC-MS. *Metabolites*, 7(1):1.

Zerbe, R.L., Miller, J.Z. & Robertson, G.L. 1991. The reproducibility and heritability of individual differences in osmoregulatory function in normal human subjects. *The Journal of laboratory and clinical medicine*, 117(1):51-59.

Zhang, T., Sun, J., Du, H., Su, H., Zhang, Y. & Jin, Q. 2018. Metabolic characterization of plasma samples in HIV-1-infected individuals. *Future microbiology*, 13(09):985-996.

APPENDIX 1: CYTOKINE PROFILE OF PLASMA STARTIFIED ACCORDING TO CLINICAL AND IMMUNOGENETICS FACTORS LINKED TO HIV PROGRESSION

Principle

Flow cytometry is a spectrometric technique used to analyse cells. Cells align in a single cell stream surrounded by sheath fluid. Light is scattered off a single cell passing in front of a laser in a laminar flow column. Light not absorbed but scattered by the cell is measured to inform on the size of the cell. Measurements of the light scattered to the sides (side scatter) informs on the optical density (granularity/complexity) of the cell. Measured forward and side scatter increase proportionally to the size and granularity of the cells, respectively. When coupled with targeted fluorescent labels, another dimension of measurements is possible. The fluorescent labels bound to their targets are excited by a laser of a specific wavelength. The excited labels emit light of a different wavelength which can be detected. The amount of light emitted is proportional to the number of fluorescent labels.

Stewart and Steinkamp (1982) first used fluorescent microspheres in addition to forward and side scatter to target and count specific cells in blood. The principle used fluorescent microbeads coated with cell-specific antibodies. Later, Lisi *et al.* (1982) described the first use of a sandwich assay to detect and quantify human immunoglobulin G (IgG). Since then, microbeads' popularity has increased among researchers. Although initially used for cell counting, microbeads are now popularly used to measure the concentrations of various soluble proteins such as cytokines.

Multiplex assays are common these days and go by the names; cytometric bead arrays (CBA) and/or Luminex® Multiplex Assays, respectively. These assays allows for multiple secreted targets to be measured concurrently through flow cytometric detection. The multiplex assays coupled to flow cytometry analysis can therefore detect and quantify as many targets as the analytical platform can distinguish by fluorescence. In fact, modern flow cytometers can measure up to 16 different colours (Perfetto *et al.,* 2004). By coating microspheres called beads with different intensities of a fluorochrome, the beads are distinguishable by the amount of light emitted. Each intensity bead is coated with a specific antibody, therefore allowing the antibody targets to be distinguished based on the bead fluorescence intensity. Secondary antibodies labelled with a different colour fluorochrome binds to the proteins on the bead in a sandwich ELISA configuration. The use of two different fluorochromes in sandwich configuration distinguishes different targets by the bead fluorescence level and determines the concentration of the target by the signal of the fluorochrome-conjugated secondary antibody.

Method

The cytokine profile of the 96 plasma samples that were subjected to metabolomics analysis was investigated. The CBA Human Th1/Th2/Th17 cytokine kit from BD Biosciences, California, USA was used to prepare the plasma samples for analysis while the BD Accuri™ C6 cytometer detected and quantified IL-2, IL-4, IL-6, IL-10, TNF, IFN- γ, and IL-17A in the respective samples.

Calibration and quality assurance preceded the analysis of the samples. The analysis of standard beads coated with different colours and intensities of fluorescence was used to perform calibration, and quality assurance on the instrument namely; the peaks and CVs of both the 6 and 8 peak beads (BD Biosciences, catalogue nr BD/653160) was recorded to ensure optimal functioning of laser.

For analysis, the plasma samples were thawed, and 25µl aliquots diluted with 25µl assay diluent. The cytokine standards were prepared and serially diluted from 5000 pg/ml to 0 pg/ml. Equal volumes of the respective capture beads (one per cytokine/interleukin) labelled with different intensities of allophycocyanin (APC) and coated with antibodies for the specific cytokines to be measured, were pooled. Of the pooled capture antibody solution50µl was added to each sample as well as standard to be analysed. Fifty microlitres of phycoerythrin (PE) conjugated secondary antibody was added to each sample and standard. The mixture was briefly vortexed to ensure mixing of the sample, beads and secondary antibodies. Incubation of the mixture in the dark for 3 hours allows the cytokines to bind to their respective beads and the secondary antibodies. Residual plasma constituents were washed off by adding 1 ml wash buffer to each sample and standard and pelleting the bead complexes. The supernatant was discarded and the beads resuspended in 300µl of wash buffer. Two thousand events (based on forward and side scatter) were recorded for each sample and standard

The FCAP Array software (BD Biosciences, San Jose, CA, USA) was used to process the data. The APC fluorescence intensities distinguished the beads while the PE fluorescence intensity was proportional to the target analyte's concentration. The data was filtered for debris by gating around the bead population on the forward and side scatter plots. Plots of the mean fluorescence intensities (MFI) of PE were drawn against standards concentrations of each bead intensity. A 5-parameter logistic standard curve was fitted for each bead intensity representing the different cytokines. The samples' PE MFI for each bead intensity was "read-off" the standard curve to obtain the concentration of the respective cytokines with the dilution of the samples factored into the calculation/algorithm. The data was exported from the FCAP Array software and further pre-processed in Excel. The concentrations of the seven cytokines for each sample was drawn up from the exported file. The samples were divided into the three sub-studies corresponding to those used in the metabolomics approach i.e. the CD4-high vs CD-low, non-progressor vs progressor and protective vs non-protective *HLA-B* groups, respectively

The data comprised of many zero values implying that the cytokine levels were most likely below the detection limit of the instrument. Crosstabs (SPSS version 26) used to investigate the pattern of the zero values within the dataset so as to determine if the pattern of the zeros differed between our test groups. No significant pattern was observed. The pattern of zeros was therefore not related to the groups of either of the sub-studies.

Subsequently, two datasets were generated for each sub-study i.e. one dataset comprising the zero values replaced by a random value between half the minimum observed value and zero while the second dataset was devoid of samples which had missing values. Upon subjecting each of the datasets to the Mann-Whitney test and calculating effect sizes, respectively (SPSS version 266), the dataset for sub-study 2, devoid of missing values yielded biologically relevant information. Upon comparing the non-progressor and progressor groups of sub-study 2, the progressor group showed an increase in IL-6 (p=0.055; ES = 0.49), IL-4 (p=0.003;ES = 0.7) and IL-2 (p=0.012; ES=0.6) levels. A logistic regression model of these three significant cytokines was able to classify 81.3% of cases of sub-study 2 correctly with an area under the curve of 0.952 with a CV of 0.79.