DEPARTMENT OF BIOCHEMISTRY
FACULTY OF SCIENCE

Computational approaches to find transcriptomic and epigenomic signatures of latent TB in HIV patients

By
Zinhle Seleka

A dissertation submitted in partial fulfilment for the Degree of
Master of Science in Biochemistry

UNIVERSITY OF JOHANNESBURG

January 2022

Supervisor: Dr Gerrit Koorsen

ACKNOWLEDGEMENTS

To my supervisor, Dr Gerrit Koorsen, who meticulously and with great care and attention to detail advised and motivated me throughout my master's studies. I would like to express my profound gratitude for his exemplary guidance, constant feedback and passion for research, which I hope to carry in my future academic endeavours.

To my parents whom I will forever be grateful for allowing me the opportunities and experiences that have made me who I am. They selflessly encouraged and supported me throughout my life. I dedicate this milestone to them.

In loving memory of my father. He will not get to see me complete this journey, but he taught me so much to get here.

SUMMARY

HIV infection promotes the progression of latent infection of *Mtb* to the active disease with the primary challenge of diagnosis being the development of efficient and sensitive methods to detect latent TB in HIV infected individuals. Previous studies have identified transcriptional signatures for active TB along with signatures predicting the risk of active TB disease in latent TB infected individuals or those with other diseases. Existing studies have also identified characteristic genes for active TB in HIV infected patients. However, no studies have identified predictive transcriptional signatures that discriminate latent TB from active TB disease in HIV positive persons as well epigenetic mechanisms associated with latent TB/HIV coinfection. The aim of this study was to develop a computational pipeline using statistical modelling and machine learning (ML) methods to identify a transcriptomic signature associated with latent TB in HIV positive patients and to identify candidate epigenetic modifications for future studies.

A novel pipeline, that leverages statistical differential expression analyses (OPLS-DA) and supervised ML and feature selection methods, was applied to an existing transcriptomic dataset (NCBI GEO repository accession number GSE37250) and the outcome of the two methodologies were integrated to define a gene signature characterising the progression of latent to active TB in HIV infected patients. Enrichment analysis was performed on the transcriptomic panel of genes to predict candidate epigenetic marks in latent TB/HIV coinfection.

An 11-gene minimal signature was identified of which the expression levels discriminate between latent TB and active TB in HIV positive patients. A broader analysis of DEGs identified by the ML and OPLS-DA revealed enrichment of pathways related to T- and B-cell receptor signalling, metabolic processes, insulin signalling, endocrine resistance and ATP-binding. Candidate epigenetic alterations associated with latent TB in the HIV positive cohort were identified using transcription factor (TF), histone modification (HM) and miRNA enrichment analyses.

This novel integrative approach to identify a discriminative latent TB gene signature provided new insights into the response mechanism of HIV co-infection with *Mtb*, and pathways that merit further investigation was identified. The genes of interest identified may provide novel diagnostic and therapeutic targets for latent TB in patients who are HIV positive.

ABBREVIATIONS

Abbreviation	Description
ANOVA	Analysis of Variance
APC	Antigen presenting cell
AUC	Area under ROC curve
BCG	Bacille Calmette-Guerin
CGI	CpG island
DEGs	Differentially expressed genes
FC	Fold change
FDR	False discovery rate
GEO	Gene Expression Omnibus
GO	Gene ontology
HIV	Human immunodeficiency virus
HM	Histone modification
IgG	Immunoglobulin
IGRA	Interferon Gamma Release Assay
KEGG	Kyoto Encyclopedia of Genes and Genomes
LTB	Latent tuberculosis
MDI	Mean Decrease in Impurity
MF	Molecular function category of the gene ontology
ML	Machine learning
miRNA	MicroRNA
MTB	Mycobacterium tuberculosis
NCBI	National Center of Biotechnology Information
OD	Other diseases
OPLS-DA	Orthogonal Projections to Latent Structures Discriminant Analysis
PC	Principal component
PCA	Principal component analysis
PLHIV	People living with HIV
ROC	Receiver operating characteristic
SNP	Single nucleotide polymorphism
TB	Tuberculosis
TCR	T cell receptor
TF	Transcription factor
TFBS	Transcription factor binding site
TSS	Transcription start site
TST	Tuberculin skin test

LIST OF FIGURES

LIST OF TABLES

TABLE OF CONTENTS

1 GENERAL INTRODUCTION

1.1 Epidemiology

Tuberculosis (TB) is a communicable disease, which according to the Global Tuberculosis Report Tuberculosis (TB) is one of the top ten causes of death and one of the leading causes of death from a single infectious agent, above HIV/AIDS [1]. In 2019, an estimated 10 million people worldwide fell ill with TB and there were a total of 1.4 million TB related deaths [1]

The World Health Organization (WHO) estimates that 1.8 billion people (almost one-quarter of the world's population are infected with *Mycobacterium tuberculosis* (*Mtb*), the pathogen that causes TB [2]. However not everyone infected with *Mtb* becomes sick as a result two-TB related conditions exist, namely latent TB infection and active TB disease. Latent TB infection usually occurs following exposure to the bacteria and is described as the state in which persons are infected with *Mtb* without any clinical symptoms, microbiological evidence or radiological abnormality [3]. An individual with latent TB infection is not considered contagious but the bacteria remain dormant in their lungs, and if left untreated latent TB can progress to the active disease state [4]. Active TB disease is the condition when an individual has signs and symptoms of TB infection and can be caused by either primary infection or reactivation of latent TB [6]. Primary TB occurs when the immune system is unable to defend itself against *Mtb* infection [6]. While reactivation TB is the reactivation of contained *Mtb* infection and is the most common form of active TB, representing 90% of cases [6]. People infected with *Mtb* have a 5-10% risk of developing active TB from latent TB during their lifetime while the remainder can contain the infection unless immunosuppressed with coinfecting viruses such as HIV. [1]

WHO reports that 1.2 million TB deaths were among HIV-negative patients and an additional 208 000 among HIV-positive patients in 2019 [7]. People living with HIV (PLHIV) are 20 times more likely to fall ill with TB, thus TB remains the leading cause of death in PLHIV [8]. The African continent has 74% of the 1.2 million HIV/TB cases worldwide [9]. Sub-Saharan Africa in particular has a significant increase in TB due to the burden of HIV [9]. South Africa, Lesotho, Botswana, Zimbabwe and Swaziland, are the southern African countries whose annual TB incidence and caseloads have increased immensely over the past 20 years, shifting what was previously a TB

problem to a crisis [10]. The latest data from Statista (2017) showed that in South Africa – where a quarter of all global HIV/TB coinfections occur – the leading cause of death was TB [11]. Epidemiologically, countries with a high burden of TB parallels the HIV pandemic.

Although present globally, the epidemiology of TB significantly varies depending on the region. South Africa in particular is one of the 30 high burden TB countries contributing to approximately 87% of TB cases worldwide [12]. Among these countries it is amongst the 14 countries with the highest burden of TB, TB/HIV coinfection and multi-drug resistant TB (MDR-TB) [12]. South Africa's TB epidemic is further driven by various factors such as low socioeconomic status, high HIV coinfection burden, delayed health-seeking behaviour in TB infected individuals, and a high instance of undiagnosed disease in communities [12]. Furthermore, the country also has the highest HIV epidemic in the world, with an estimated 7.7 million PLHIV in 2018 and accounts for a third of all new HIV infections in Sub-Saharan Africa [13].

Bacterial, host, and environmental factors influence the progression of latent TB to active TB [14]. TB is an infectious disease that begins in the lungs through infection via aerosol droplets or the bloodstream [15]. Once the *Mtb* pathogen has entered the bloodstream, the bacteria spreads through the body and result in the infection of various tissues; however with latent TB the bacteria stay dormant before resulting in active TB [15]. In more than 90% of cases immune responses which are cell-mediated, control progression of the disease, result in latent TB [16]. While less than 10% of these cases advance to active TB causing approximately 3 million deaths globally every year [16].

In HIV infected individuals, the presence of other infections such as TB accelerates the rate of HIV replication [17]. This acceleration results in greater levels of infection and consequently a quicker progression to the AIDS stage [17]. HIV promotes the progression of latent infections of *Mtb* to the active disease and consequently increases the rate of occurrence of TB shortly after HIV infection [18]. In the past, the main challenge of diagnosis has been the development of efficient and sensitive methods to detect latent TB in HIV-infected individuals [15]. However, in more recent years, the significant progress made in genomics has provided an important reference to assist in the diagnosis of HIV combined TB infection [19].

1.2 Pathogenesis and Mechanism of Latent TB/HIV Coinfection

The pathogenesis of *Mtb* is complex, involving an elaborate interaction with the host [20]. Key factors such as the ability of *Mtb* to survive in macrophages, its preference for the lung, the formation of granulomas and its long term persistence remain poorly understood [20]. *Mtb* infect via the respiratory tract where they invade and replicate within the alveolar macrophages [21]. The macrophages then illicit an immune response resulting in the formulation of a granuloma, which is a structure composed of macrophages, dendritic cells, lymphocytes, neutrophils and sometimes fibroblasts [21]. This structure acts to contain the bacilli and limit *Mtb* replication; however, just the simple formation of a granuloma is not sufficient enough to control infection as a person with active TB may contain multiple granulomas in the lungs [21]. Instead, the granuloma requires optimal immunological function in order to contain the bacilli [21].

This ongoing immune response against *Mtb* has been shown to increase the replication of HIV-1 in the blood and at the sites of *Mtb* infection in the lungs [22] as seen in Figure 1.1. HIV causes a depletion of CD4 T-cells, which contributes to the progression of latent TB to active TB [21]. HIV also affects other cells such as macrophages, and thus influences cytokine production, which can also prevent a host from containing latent *Mtb* infection [21]. As with all opportunistic illnesses in HIV-infected, the risk of TB increases at lower CD4 counts.

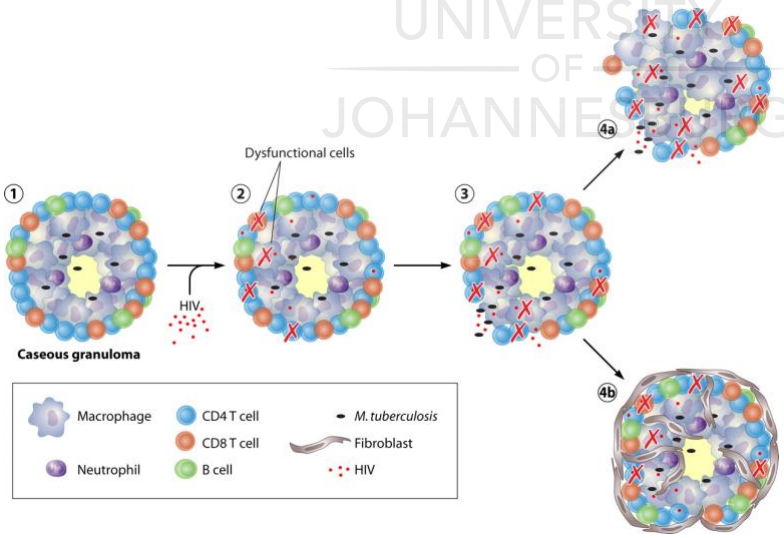


Figure 1.1 Proposed mechanism of HIV induced reactivation of latent TB.
Stage 1 illustrates a necrotic granuloma functioning normally in a latent TB infected individual. This is proceeded by Stage 2 whereby HIV enters the granuloma causing functional changes within macrophages and T cells and also kills activated T cells. Stage 3 results in the decrease in a number of T cells and an increase in cellular dysfunction resulting in the functional disruption of the granuloma, which may ensue increased dissemination. Stage 4a indicates the disruption of granulomas shortly after HIV infection causing continued *Mtb* dissemination as well as early TB reactivation. Stage 4b demonstrates the fibrotic granulomas temporarily re-establishing granuloma containment, preventing reactivation [21].

Although several hypotheses regarding the exacerbation of TB by HIV have been put forward, the mechanisms by which HIV infection disrupts TB granuloma function resulting in increased morbidity and mortality are poorly understood. Thus, HIV/TB coinfection presents enormous diagnostic and therapeutic challenges on immensely burdened countries that are heavily infected [18].

Considering most patients infected *Mtb* do not develop the active clinical disease, latent TB exhibits a unique and challenging opportunity to comprehend the intricate relationship between host cells in the granuloma [23]. Another major contributor to this problem is that no clear pattern of host responses are linked with latent TB infection since molecular correlates of latent TB infection have been difficult to identify [5].

1.3 Diagnosis of Latent and Active TB

The standard diagnostic test for TB is the Tuberculin Skin Test (TST) which involves intradermal injection of a purified protein derivative (PPD) – a mixture of more than 200 antigens which are shared by other mycobacteria [24]. Injection of this PPD leads to a delayed hypersensitivity response and thus causing a cutaneous induration at the site of injection, usually between 48-72 hours [24]. TST has several limitations with regards to its specificity and sensitivity, as it may be positive in patients that have a prior Bacille Calmette-Guerin (BCG) vaccination to those that have had exposure to non-tuberculous mycobacteria [24]. In addition, false negative TST results can frequently occur in patients that have impaired T-cell function [24].

Another assay that was introduced for the detection of *Mtb* infected individuals is the Interferon Gamma Release Assay (IGRA) which is a blood test that works by measuring interferon-gamma release in response to T-cell stimulation cause by specific *Mtb* antigens [24]. IGRAs have progressively replaced TST as they are meant to offer improved specificity and sensitivity [25]. However, IGRAs major limitations are that they are unable to discriminate between active and latent TB [26] and are therefore insufficient for marking the disease status.

Although there have been recent advances in the identification of TB biomarkers – specifically those that have the potential to diagnose and differentiate active TB from latent TB [27]; the problem lies when an individual has an immunosuppressing infectious disease such as HIV. The sensitivity of TB diagnosis is further limited by the presence of HIV infection as it results in the increase of false negatives [28]. There is

only limited data available that describes IGRA performance in HIV-infected individuals in which their immunological impairment could affect the performance of this lymphocyte based assay [26]

Due to their high risk of TB, PLHIV need to be regularly screened for TB [29]. This screening involves investigating the presence of a current cough, fever, weight loss and night sweats; the presence of one of more of these symptoms would then prompt the application of a TB diagnostic procedure as the individual would be presumed to have TB [29]. To diagnose latent TB in PLHIV, the TST should be used. IGRA might be used instead of TST in settings where BCG vaccination coverage is high and whereby organisation of testing is feasible. A positive TST or IGRA result among PLHIV without any indicating signs of active TB should be considered to indicate latent TB in an individual and this requires the implementation of TB preventive treatment [29].

New tests to diagnose latent TB are required for immunocompromised individuals such as HIV infected persons. Ideally, these tests should have numerous characteristics such as high sensitivity for all populations that are at risk including high specificity, reliability, stability over time and objective criteria for a positive result [26]. And because latent TB occurs at such a high incidence, the tests should be inexpensive and easy to perform [26]. Therefore novel biomarkers that target the host immune responses against *Mtb* may aid in improving clinical tests [16].

1.4 Treatment for Latent TB Infection in PLHIV

Prevention of active TB by means of latent TB treatment is one of the greatest strategic elements to eliminate TB. Clinical trials have illustrated that latent TB treatment in PLHIV reduces the risk of active TB, particularly in those with a positive tuberculin skin test [30]. There are several latent TB treatment options available for PLHIV [30]. Until recently, isoniazid (INH) preventive therapy has been the most widely recommended and used regimen for the treatment of latent TB infection in PLHIV [30]. However, completion of INH therapy in both HIV infected and uninfected persons is very low [31]. The WHO recommends screening HIV infected individuals for a cough, fever, night sweats and weight loss [31]. If no symptoms are present INH preventive therapy is still recommended even in the absence of skin testing [31]. The main challenge to this strategy is the lack of proven efficacy of INH in tuberculin skin test negative persons [31]. Thus, it is imperative to identify other safe and effective treatment regimens with high completion rates.

For persons with positive tuberculin skin tests, the Centers for Disease Control and Prevention (CDC) recommends that healthcare providers prescribe a short-course regimen when possible since patients are more likely to complete a shorter treatment plan [32]. The latest CDC recommended treatment for individuals with latent TB and HIV taking antiretroviral medications (with acceptable drug-drug interactions) is 12 weeks of isoniazid and rifapentine antibiotics [32]. Another treatment option is four months of daily rifampin; however, this regimen should not be administered in HIV infected persons taking some combinations of antiretroviral therapy (ART) [30, 32]. An alternative treatment for individuals taking antiretroviral medications with significant drug interactions is nine months of daily isoniazid [32].

1.5 Transcriptome Profiling for Disease Prediction

Over the last few decades, transcriptome profiling has been one of the most widely used approaches to analyse human diseases at a molecular level [33]. Many molecular biomarkers and therapeutic targets have been identified for various human pathologies through gene expression studies [33]. The transcriptome contains all sets of RNA transcripts of the genome in a specific tissue or cell type [33]. The two key techniques in the field of transcriptomics include microarrays and RNA sequencing (RNA-Seq) [34]. Microarrays quantify a set of predetermined transcripts/genes through hybridisation, while RNA sequencing utilises high throughput sequencing in order to capture all sequences of the whole transcriptome [35]. Transcriptomic analysis has enabled the study of gene expression changes in different organisms, which is fundamental to understanding human disease [34]. Measuring the expression of an organisms' genes in different conditions provides information of how the genes are regulated [34].

1.5.1 Microarray Gene Expression and Gene Expression Signatures

Microarray technology is used to measure the expression levels of thousands of genes simultaneously, during a single experiment [36]. The primary use of DNA microarrays is transcriptional profiling. Microarrays enables the identification of differentially expressed genes (DEGs) between two or more biological conditions. A gene is described as differentially expressed when an observed difference or a change in expression levels or read counts between two experimental conditions is statistically significant [37]. A microarray is a collection of DNA probes bound to a fixed surface in such a way that the identity of each probe can be determined through its position on

the array. The probes of microarrays comprise of a string of nucleotides that are complementary to the sequence of the gene being investigated [38]. The probes can be oligonucleotides, complementary DNA (cDNA) or small fragments of PCR products corresponding to mRNAs [38]. Typically an oligonucleotide probe is single-stranded and between 25 and 70 bases long [38]. While cDNA probes are double-stranded with the length of a full gene product (2kb on average) [38].

Omics repositories such as the National Center of Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) and EMBL-EBI ArrayExpress accumulate and provide gene expression data from thousands of studies facilitating the reanalysis of gene expression data by researchers [39]. The generation of large datasets from microarray experiments, that can be accessed through such repositories, presents unique challenges in acquiring, annotating, analysing and warehousing that data [38]. More recently, research requires much more from microarray experiments than just a list of up or downregulated genes, the functional associations between those genes must also be made [38]. Gene annotation requires the assimilation of functional information about protein products and motifs [38]. Gene Ontologies (GO) are used to capture and describe these annotations and provide biological classifications, specific molecular interactions as well as subcellular localisations [38].

1.5.2 Transcriptomic Signatures for Diagnosing Tuberculosis

A means of screening populations that have latent infection or are at high-risk of developing active TB disease is crucial in applying prophylactic therapy for the prevention of TB [40]. Transcriptomic signatures for individuals with active TB have been proposed and indicate a promising diagnostic tool [41]. More recent studies have explored the identification of transcriptomic signatures which differentiate individuals with latent TB infection from those with active TB [41-44]. Although biomarkers have been identified they have not yielded translatable outcomes for understanding protective immunity in the host. Thus, our understanding of the relationships that exist between the states of latent TB and active TB along with the immune factors that affect possible transition between states is very limited [40]. Moreover, the overlap of pathology caused by *Mtb* and the HIV virus presents further challenges. TB/HIV coinfection increases the risk of active TB, however, the determinants of this synergism are uncertain [45]. As such, transcriptomic data remains insubstantial for differentiating latent TB from active TB in HIV infected persons.

1.6 Identification and Classification of Gene Expression Signatures using Computational Biology

The rapid developments in genomics, through the applications of technologies such as microarrays or high-throughput performance sequencing, have produced a significant amount of biological data [46]. This has presented new opportunities and challenges in the fields of computational biology and bioinformatics since this large data needs to be analysed to draw conclusions from this data [46]. Increasingly, sophisticated computational methods and tools, such as machine learning tools or gene expression analysis tools, are required and applied for downstream analyses of complex biological data. Multivariate statistical methods, particularly for both supervised and unsupervised clustering, principal component analysis, regression and visualization tools, such as heatmaps, have proved vital to the interpretation of outcomes for these complex data which are generated by combinations of epigenetic changes and molecular events [47].

The use of numerous computational approaches and tools in order to comprehend transcriptional and epigenetic changes and their impact on disease has thus become increasingly more important for data generation, mapping and management as well gene analysis and therapy [47]. Computational tools play a critical role in forming testable hypotheses and directing the selection of key experiments through the analysis of complex and large genomic information, which is not attainable using only traditional approaches [48]. Managing and mapping for "-omics" studies are important steps, but the real challenge now is the interpretation of this data in order to quantify risk and drive therapeutic development and disease management. [47].

A gene signature or a gene expression signature (GES) is a combined set of genes with a unique characteristic pattern of altered gene expression that can distinguish patients with different conditions [36]. A GES can be used for the purposes of diagnosis, prognosis or the prediction of therapeutic response [36]. GESs are typically determined through the use of statistical methods such as fold change and the *t*-test [36], but over the more recent years, machine learning has become the method of choice for identifying GESs. With these recent methods, similar GESs that are most likely induced by the same or related biological mechanisms can be established [49]. Identifying genes that are positively or negatively correlated can provide insights into previously unknown connections among biological networks [49]. For instance, distinct diseases may result in overlapping mRNA expression patterns emerging from the

same or related immune response processes [49]. Moreover, mutations that are induced by similar GESs may enable functional associations with biological processes even if the affected genes do not share sequence similarities. Likewise, drugs employed for different therapeutic applications may share similar GESs owing to related mode of actions [49]. GES technology may be extremely beneficial for identifying novel drug targets leading to the discovery of novel pharmaceutical treatments for various diseases and for addressing key biological and human health research problems [49].

1.6.1 Statistical Techniques that can be applied to Microarray Datasets

Statistical analysis is defined as the study of the collection, organisation, interpretation, analysis and organisation of data and involves the drawing of inferences from the samples to the whole population [50]. Statistical analysis usually requires the proper design applied to the study, a suitable selection of the study sample as well as the choice of a suitable statistical test [50]. Statistical analysis is one of the most important techniques used to identify transcripts whose expressions change significantly across different samples or experimental conditions [51]. The complex nature of gene expression studies presents challenging statistical problems that require the use of a number of specialised statistical techniques to exploit each data set [52]. Statistical analysis may be represented by both univariate and multivariate data analysis [50]. Univariate analysis is the simplest of statistical analyses where only one variable is analysed [53]. Multivariate data analysis, which is used to compare more than two variables, is beneficial for summarising data and examining patterns in multidimensional data [53]. Before utilising any statistical test, it is necessary to apply some pre-filtering processes, normality tests and normalisation, respectively to obtain results with reduced error [51]. The Fold Change (FC) is a fundamental method that is widely used to identify DEGs [54]. The FC is calculated as a ratio of averages from control and test sample values initially used [54]. Cut-offs are used and genes with FC values below or above thresholds are selected [54]. Other statistical methods are also used, which apply three steps. First, a statistical test (for e.g. Student's *t*-test) is performed to derive the *p*-value for each gene in the microarray study [54]. Second, the p-values are compared to a selected threshold that has an acceptable False Discovery Rate (FDR), after which a list of genes is attained [54]. Third, using the FC level thresholds, up and down-regulated genes from the above list are selected [54].

Statistical testing procedures can be broadly classified as parametric and nonparametric tests [55]. The most common statistical procedures are parametric

statistics, which make certain assumptions about the distribution of the sample population such as that the sample distribution has the same parameters (means, standard deviations) and are the same shape (normally distributed) [55, 56]. If the data does not meet these assumptions then one possible alternative is to use nonparametric statistics that make no such assumptions about the parameters or shape of the population distribution [55, 56]. Different parametric and nonparametric testing methodologies can be used for discovering DEGs from microarray datasets [51]. The types of popular parametric tests which can be used to identify DEGs include: *t*-test, ANOVA 1 test and Pearson's correlation test. While common nonparametric tests that can be applied include: Permutated t-test, Wilcoxon Ranksum Test, Significance Analysis of Microarray, Linear Models for Microarray Data (Limma) and Shrink-t [51].

1.6.2 Machine Learning Classification and Feature Selection

In biological research, observation data is used to derive classes into which patients can be assigned to. The classes could include patients with a disease subtype, disease state or patients responding to a particular treatment. Consequently, these classes may be used for personalised healthcare rather than making use of a one-size-fits-all approach. The problem is usually not knowing in advance which classes are relevant for each patient; hence the main challenge lies in defining criteria to assign individual patients to known classes [57]. The classification methods used in microarray studies are distinct in the manner that they deal with the underlying intricacy of the data and in the technique employed to build the classification model [58].

Microarray DNA technology is computationally expensive and generates high-dimensional data with minimal sample size [59]. This high dimensionality is largely owed to the vast number of values generated for every gene in a genome [59]. Employing such a high dimensional dataset can result in over-fitting of the model [60]. To address this issue the dimension of the data needs to be reduced by a considerable amount [60]. In the past few years, Machine Learning (ML), a subset of Artificial Intelligence has gained traction in the field of genomic studies [60]. The primary purpose of ML is to enable a model to train and learn a dataset to make predictions or decisions in the future without being explicitly programmed to do so [60].

ML is generally categorised as Supervised, Unsupervised and Reinforcement learning [60, 61]. Supervised learning involves feeding a learning algorithm labelled data and over time the algorithm learns to approximate the relationship between examples and

their labels [61]. Once fully trained, the supervised learning algorithm is able to observe a never before seen example and predict a label for it [61]. Unsupervised learning on the other hand involves no labels, instead, the algorithm is fed a lot of data and provided with the tools required to understand the properties of the data. The algorithm can then learn to group, cluster and/or organise the data in a way that a human or other intelligent algorithm can make sense of the new organised data [61]. Reinforcement learning is a feedback-based ML technique where agents are trained on a reward and punishment mechanism [61]. The agent is rewarded on desired behaviours or moves and/or punishing undesired ones [61]. In doing so the agent attempts to minimise incorrect moves and maximise the correct ones [61].

Microarray data classification is typically performed in two phases: (i) Feature Selection which selects the most relevant features from a large dataset in order to reduce noise, overfitting and computational overheads. (ii) Classifier Training which builds a model from selected features to accurately and reliably classify a given microarray sample [62]. Feature selection is a dimensionality reduction technique and assists in preserving informative attributes [63] and is significantly useful in fields with datasets containing too many features and relatively scarce samples, such as DNA microarray [63]. Feature selection techniques assist in overcoming model overfitting, handling the high dimension, maximising prediction accuracy and model training time [64, 65]. The feature selection outcome is usually the optimal number of features relevant to a given class label, which ultimately contributes to the process of prediction [64, 65]. The primary purpose of applying feature selection to gene expression data is to select the most regulating genes and discount redundant genes that do not contribute to the target class [66].

1.6.3 Pathway and Network Analysis of Differentially Expressed Genes

In the past, gene expression data were analysed on a gene-by-gene basis, disregarding complex interactions and association mechanisms, thus overlooking the presence of important underlying biological signals [67]. Researchers began to realise the concerted manner in which genes act and that cellular processes are a result of the complex interactions between different genes and molecules [68]. Genes can be categorised based on various traits such as sequence, function and interactions [67]. Grouping genes by biological pathway is usually the most relevant approach [67]. Pathway analysis is defined as an approach that identifies differential expression patterns in a dataset by considering pathways structure. In pathway analyses,

researchers are typically interested in determining pathways associated with a biological condition and establishing crucial components of those pathways which explain the association [67]. Numerous repositories exist containing information about thousands of biological pathways which contain up to several hundred proteins [67].

High-throughput technologies such as microarrays generate large gene lists of interest as their final outputs, which can range in size from hundreds to thousands of genes [68]. Over the decades, bioinformatics methods have made use of biological knowledge generated in public databases such as Gene Ontology (GO) to perform network enrichment analyses that This allows researchers to systematically dissect these large gene lists, ascertain which genes are differentially expressed (enriched) and have an association with disease phenotypes [68, 69] and gain information on the relationships between genes that are provided by gene networks [68]. In order to gain insight into the biological significance of the alterations in gene expression levels of DEGs, researchers utilise the GO enrichment analysis to establish whether GO terms relating to specific molecular functions, biological processes, or cellular components are over or under-represented within a gene set of interest [70].

1.7 Epigenetic Mechanisms

Epigenetics has emerged as an important field in studying the influence of non-gene factors on the traits and functions of organisms [48]. The term ‘epigenetics’ was coined by biologist Conrad Waddington in 1942, by using studies of how environmental influences develop in conjunction with genotype, leading to the development of phenotype [71]. The field of study involves changes in gene expression caused by factors other than an individual's DNA sequence. Every cell type has a unique epigenome which allows for specific cell differentiation and since a single genotype can be associated with several phenotypes, it is believed that infinite epigenomes exist for a single genome sequence [71]. Epigenetic changes refer to changes in DNA structure as a result of DNA modification after replication or post-translational modification of proteins that are associated with DNA [72]. Contrary to mutations, epigenetic changes occur very rapidly and are reversible [72]. The major epigenetic mechanisms include DNA methylation, histone modifications, and non-coding RNA-associated gene silencing [48, 72] and an eminent goal for this area is to determine regions in the genome that are susceptible to these epigenetic modifications.

1.7.1 DNA Methylation

A key driver in epigenetics is DNA methylation, which is a biological process resulting in the addition of a methyl group to the fifth carbon of a cytosine residue in CpG dinucleotide sequences to form 5-methylcytosine (5mC) [72]. This process is catalysed by a family of DNA methyltransferases (DNMTs) [72]. The main DNMTs responsible for DNA methylation include DNMT1, DNMT3A, DNMT3B and DNMT3C [73, 74]. Methylation of DNA is the most extensively studied epigenetic mechanism. DNA methylation regulates gene expression through the recruitment of proteins involved in gene expression or by inhibiting transcription factor binding to DNA [75].

DNA methylation, occurring in the context of the CpG dinucleotide, has profound effects on gene expression by modifying the accessibility of DNA to transcription factors [72]. CpG islands (CGIs) are short DNA sequences of 200 bp to several kilobases in length usually located near the promoter, that deviate significantly from the average genomic pattern by being CpG-rich, GC-rich, and predominantly non-methylated [72, 76]. It is widely accepted that DNA methylation that occurs on CGIs acts as a silencing mechanism [77]. There are ~29 million CpGs located in the human genome, of which 60-80% are methylated and approximately 7% of CpGs are found in CGIs [78]. Additionally, ~70% of annotated genes have promoter regions containing CGIs which can be highly, partially or lightly methylated [76].

The process of demethylation is more complex and can occur through either active or passive mechanisms [79]. Active DNA methylation occurs through an enzymatic process that modifies or removes the methyl group from 5mC. Ten-eleven translocation (TET) enzymes such as TET1, TET2, and TET3 are involved in active DNA methylation through the oxidation of 5mC and promoting locus-specific removal of DNA methylation [80]. Passive DNA methylation, in contrast, refers to the loss of 5mC that can occur in the absence of functional DNA methylation maintenance machinery during successive rounds of replication [81].

Over the years, it was believed that DNA methylation played a crucial role in repressing gene expression by blocking the promoter sites where transcription factors should bind [82]. At present, the exact role of methylation in gene expression is unclear; however, it seems that DNA methylation is vital for cell differentiation and embryonic development [82]. Evidence has observed the role of methylation in gene expression mediation. Studies have shown that methylation occurring near gene promoters considerably varies depending on cell type, whereby more methylation of promoters is

correlated with little or no transcription [83]. Although overall methylation levels of particular promoters are similar in humans, there are significant disparities in specific and overall methylation levels between different cell lines and tissue types [83].

Given the integral role of DNA methylation in gene expression and cell differentiation, errors in methylation can give rise to a number of perilous consequences such as various diseases [83]. Over the years, changes in DNA methylation have been detected in many human diseases [78] such as cancer, autoimmune diseases, neurological disorders, metabolic disorders, and a range of birth defects caused by defective imprinting mechanisms [79, 84]. To date, a significant amount of DNA methylation research has focused on cancer and tumour suppressor genes [83]. Tumour suppressor genes are usually silenced in cancer cells as a consequence of hypermethylation, which represses transcription of the promoter regions [83, 85]. Comparatively, cancer cell genomes have shown to be hypomethylated in comparison to normal cells [84]. DNA hypermethylation has been more extensively studied as opposed to DNA hypomethylation as a cause of oncogenesis [85]

1.7.2 Histone Modification

Epigenetics is not just limited to the study of DNA, the post-translational modification on histone proteins play a significant role in gene expression regulation [86]. Histone modification modulates the structure of the chromatin, thus limiting the accessibility of DNA [87]. Several types of histone modifications exist, with methylation, acetylation, phosphorylation and ubiquitination being the best studied and most important with regards to the regulation of chromatin structure and transcriptional activity [88]. The main histone-modifying writer enzymes (i.e enzymes adding marks on histones) are histone methyltransferases (HMTs), histone acetyltransferases (HATs), protein kinases (PTKs) and ubiquitin ligases [87]. While their respective eraser enzymes include histone deacetylases (HDACs), histone demethylases (HDMs), protein phosphatases (PPs) and deubiquitinating enzymes (DUBs). In addition to epigenetic writers and erasers, there are also epigenetic readers which are molecules that are able to recognise and bind to epigenetic marks created by writers, thus establishing their functional consequences [87].

Histone acetylation, which involves the addition of an acetyl group to histones, is mediated by two groups of enzymes the HATs and HDACs [87]. HATs catalyse the transfer of an acetyl group from an acetyl-CoA molecule to an amino acid group of the

target lysine (K) residues in the histone tails resulting in the removal of a positive charge on the histone, thereby weakening the interaction between DNA and histones [89-91]. Histone methylation is facilitated by HMTs, including lysine methyltransferases (KMTs) and arginine methyltransferases (PRMPTs) as well as by HDMs. Contrary to acetylation of the histone lysine which affects the electrical charge of the histones, methylation of lysine or arginine residues affects the recruitment and binding of regulatory proteins to chromatin which in turn affects the histones interaction with DNA [92-94].

Histone phosphorylation is controlled by PTKs and PPS which have opposing modes of action [87]. Kinases add phosphate groups, while phosphatases remove phosphates [95, 96]. The three main functions of phosphorylated histones include DNA damage repair, the regulation of transcriptional activity and the control chromatin compaction linked to meiosis and mitosis [95, 96]. In contrast to histone methylation and acetylation, histone phosphorylation works in tandem with other histone modifications allowing for mutual interaction between the modifications. This crosstalk between the histone modifications may determine transcriptional outcomes such as transcriptional activation or repression [97]. Protein ubiquitination is an essential post-translational modification that controls substrate degradation and quantity and quality of various proteins thus ensuring cell homeostasis [98]. Histone ubiquitination is carried out by ubiquitin ligases and can be removed by DUBs [87]. Monoubiquitination is involved in protein translocation, DNA damage signalling as well as transcriptional regulation [87]. Polyubiquitination marks the protein for activation or degradation in certain signalling pathways [98-100]. As with histone phosphorylation, crosstalk exists between histone ubiquitination and other histone modifications [98-100].

Histone modification and DNA methylation affect one another during nucleosome remodelling and gene expression regulation, which may influence the development of cellular processes [101]. For instance, DNA methylation alone is unable to directly maintain stable gene silencing, therefore histone modifications may assist in directing DNA methylation patterns and thus provide long-term stability of gene repression [102]. In contrast, histone methylation may only cause reversible gene suppression, where DNA methylation would be a secondary event leading to stable long-term repression [103].

1.7.3 Non-coding RNA

Less than 2% of mammalian genome transcripts have a protein-coding function, while the remaining 98% are non-coding RNA (ncRNA) [104]. ncRNA are categorised into two types namely microRNA (miRNA) and long non-coding RNA (lncRNA) [104]. Studies have shown the promoter regions of miRNA and lncRNA genes contain different epigenetic modifications and involve various biological processes through the interaction with transcription factors [105-108]. Consequently, any abnormalities occurring in these transcriptional processes can result in disease [109].

MiRNAs are a class of short double-stranded RNAs, approximately 18-25 nucleotides in length [110]. This class of ncRNAs are responsible for silencing mRNA translation through direct interaction with the transcript [111]. Recent advances have established miRNAs as epigenetic modulators that affect the target mRNAs protein levels with modifying the gene sequences [112]. Furthermore, miRNAs can be regulated by epigenetic modifications such as DNA methylation, histone modifications and RNA modifications. The complementary actions of epigenetic pathways and miRNAs form a miRNA-epigenetic feedback loop which has a substantial influence on the proliferation of gene expression [112]. The misregulation of this feedback loop impedes the pathological and physiological processes, leading to various human diseases [112].

LncRNAs are RNA molecules which are longer 200 nucleotide bases that lack protein coding potential and are transcribed by the RNA polymerase II [113]. Many lncRNAs play an essential regulatory role in varied biological processes and their dysregulation can contribute to different diseases [113]. Depending on their cellular localisation and interacting molecules, lncRNAs mechanisms of action are heterogeneous in nature [113]. Different lncRNAs in the nucleus act by directing epigenetic regulators to specific loci or by coordinating chromatin folding and compartmentalisation so as to achieve enhancer-promoter communication [113]. Genome studies have highlighted that mutations occurring in regulatory regions that alter either enhancer and promoter sequences or their chromatin state, bring about abnormal expression of lncRNA in diseased tissues [114-117].

1.7.4 Transcription Factors

Gene transcriptional regulation is an integral part of tissue-specific gene expression as well as gene activity in response to external stimuli [118]. The main regulators of DNA transcription are transcription factors (TFs) [119]. TFs are proteins that bind to the

upstream regulatory elements in the promoter and enhancer regions of DNA in order to control transcription [120]. When a cell encounters changes in environment, it responds by modifying one or more TFs [121]. Through transcriptional changes, TFs are then responsible for altering cellular function and deciding the cell fate [121]. Due to their importance in cellular programming, measuring the differential TF activity in two different conditions provides critical insight, particularly when the involved cellular processes are unknown [121].

Transcription factors contribute to the epigenetic control of gene expression through several contexts [122]. In eukaryotes, TFs interact with a range of mechanisms that methylate DNA, post-transcriptionally modify histones and regulatory proteins, reorganise chromatin structures, remodel nucleosomes and recruit coregulators in order to regulate transcription [123]. These genetic and epigenetic mechanisms mark regulatory regions of a gene of interest [123]. The complex nature of individual gene regulation is driven by environmental factors and cellular requirements for the gene product 103]. As such, TFs acting at specific enhancer and promoter regions evolve to match the gene's regulatory needs [124]. The TF binding sites and the DNA sequence's propensity to assemble nucleosomes defines the promoter architecture and function [124]. However, because the DNA is weaved with several forms of epigenetic information contributing to gene regulation, the genetic and epigenetic information, which act at the same regulatory sequences, integrate and complement each other in all genes [124].

1.8 The Role of Epigenetic Modifications in TB Infection

Epigenetic mechanisms play a pivotal role in the regulation of gene expression during cellular response to extracellular stimuli [125]. Studies have shown that *Mtb* can alter the host epigenome, but the mechanism of these *Mtb*-induced epigenetic alterations is not fully understood yet [125]. Genome-wide association and candidate gene studies have revealed complex links between TB susceptibility and heritable genetic factors however their the reproducibility and consistency of these studies remains uncertain [126, 127] . In order to answer important questions regarding host susceptibility, further research on the role of epigenetics during immune response regulation in the context of TB is essential [125].

A limited number of studies have reported on the interaction between *Mtb* infection and host epigenetic machinery changes, although the precise molecular mechanisms have yet to be determined [125]. Epigenetic alterations such as DNA methylation, histone modifications and miRNA mediated up or downregulation of immune-related genes play a pivotal role in immunomodulation following *Mtb* infection (Figure 1.2) [125]. Epigenetic modification induced by *Mtb* can promote either host defense or survival of *Mtb*. [128] Thus *Mtb* may be regarded as a potential host epigenome modulator with these epigenetic changes being either beneficial or harmful to *Mtb [128]*.

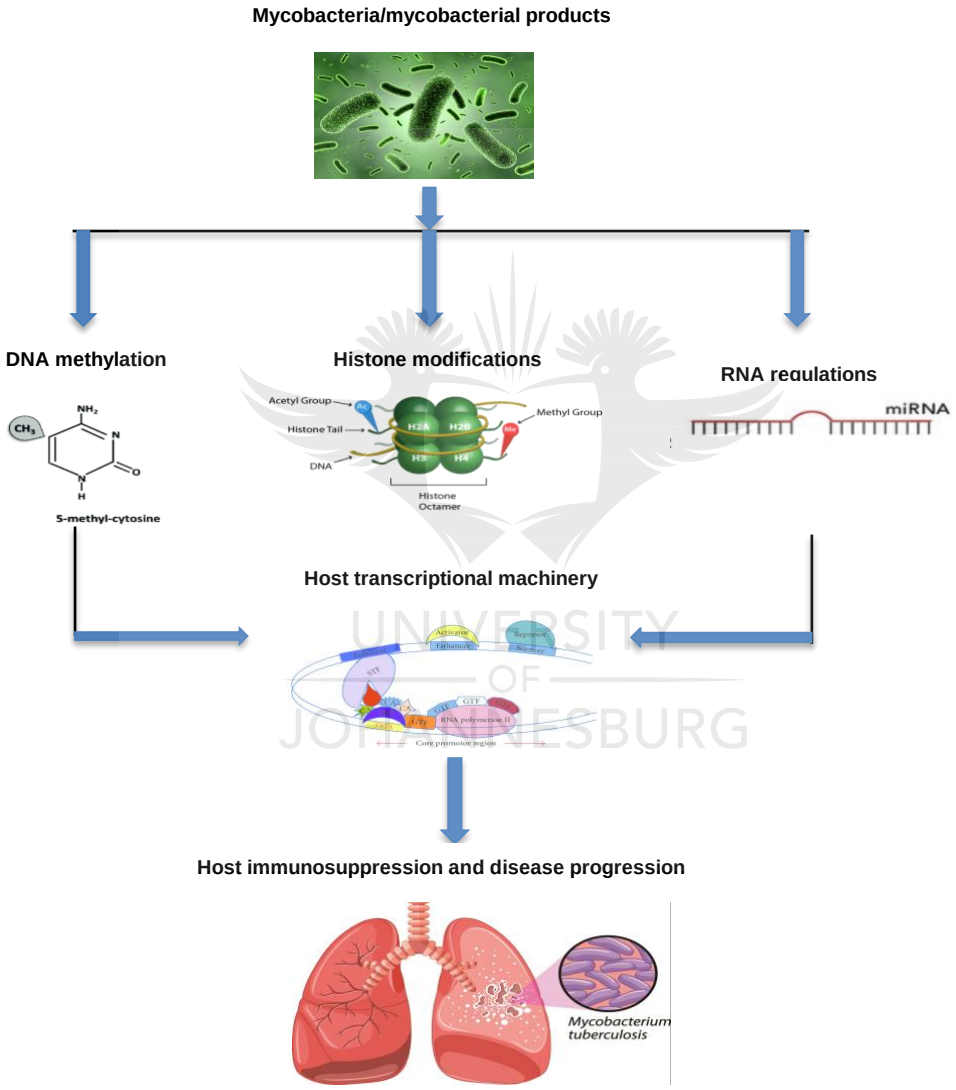


Figure 1. 2 Interaction between the host epigenetics factors and mycobacteria. Adapted from Kathirvel & Mahadevan
Epigenetic modifications such as DNA methylation, RNA mediated regulations and histone modifications are interlinked and regulate patterns of gene expression [125].

Milena *et al.* have reported on how HIV alters effector functions of monocytes [129]. Once these alterations associated with epigenetic changes are identified, these could

be used as targets in therapies that are aimed are reducing the systematic activation state in HIV infected patients [129]. Understanding epigenetic changes in HIV/latent TB coinfection is of major interest in targeting intervention and therefore providing appropriate diagnostic methods and pharmaceutical development [47]. Further studies focusing on how epigenetic factors influence and regulate the activation of active TB from latent TB infected individuals are required. To date, there appear to be no studies that have addressed the impact of epigenetic factors in latent TB patients coinfected with HIV [125].

1.9 Rationale

To the best of our knowledge, there exists no study integrating the transcriptional and epigenetic changes in the immune response genes associated with HIV positive patients coinfected with latent TB. Therefore, developing a preparative computational workflow in order to select transcriptomic and epigenomic markers in the genes of interest to establish their role in latent TB/HIV coinfection will be vital in the prevention of progression of latent TB to active TB and possibly contributing towards the development of suitable diagnostic methods.

1.10 Aim & Study Objectives

The overall aim of this study was to develop a computational pipeline using statistical modelling and machine learning methods to identify a transcriptomic-based minimal gene signature associated with latent TB in HIV positive patients and to identify candidate epigenomic markers for future studies. The study focused on an integrative transcriptomics approach and an epigenomics approach, each having its own objectives.

For the integrative transcriptomics approach to identify a latent TB gene signature in HIV infected individuals, the specific objectives are outlined as follows:

1. Apply OPLS-DA statistical modelling on a previously published transcriptomic dataset to find a list of differentially expressed genes to distinguish the latent and active TB classifications.
2. Utilise machine learning and feature selection methods on the same dataset to identify differentially expressed genes distinguishing the latent and active TB classifications.

3. Perform a network enrichment analysis to identify biological pathways related to the genes of interest using the outputs from the statistical and machine learning methods.
4. Integrate the outcomes of the statistical and machine learning approaches to define a minimal latent TB gene signature in HIV infected persons.

For the epigenomics approach to determine epigenetic mechanisms involved in latent to active TB progression in HIV infected persons, the main objective are the following:

5. Using the identified gene signature, perform transcription factor, histone modification and miRNA enrichment analyses to obtain candidate epigenetic marks in latent TB/HIV coinfection.

2 AN INTEGRATIVE TRANSCRIPTOMIC APPROACH TO IDENTIFY A MINIMAL LATENT TB GENE SIGNATURE IN HIV INFECTED INDIVIDUALS

2.1 Introduction

The development of transcriptomics approaches has enabled the discovery of genes whose differential expression is associated with disease. Transcriptional profiling has previously been used extensively to identify biomarkers in pathogen infections [130-132]. Berry *et al.* identified the first complete human blood transcriptional signature of 393 transcripts for active TB and an 86 transcript signature discriminating active TB from other infectious diseases [130]. A later study reported a 16-gene signature in whole blood, which predicted the risk of active TB in individuals with latent TB infection [133]. Researchers have also identified characteristic genes and pathways for TB in HIV infected patients [134]. However, the genes involved in the progression from latent to active TB in HIV positive patients remain unclear.

In a noteworthy study, Kassa *et al*. aimed to identify genes that have the potential to discriminate active TB from non-active TB in HIV infected patients from Ethiopia. Five genes of interest were suggested (CD8A, CCL22, FCGR1A and TNFRSF1A). Whole blood samples from 106 patients among three clinical groups (active TB/HIV positive, latent TB/HIV positive, TB negative/HIV positive) were used [135]. The latent TB group reflected patients who tested positive for the TST, while the TB negative group reflected patients who tested negative for the TST. However, since their results revealed no difference in gene expression between HIV infected patients with latent TB and those who tested negative for TB, they combined these two groups and compared this combined group (described as non-active TB) with the active TB/HIV positive patients. Since the study combined latent TB and TB negative groups, it is therefore still unclear which genes are differentially expressed in HIV positive patients when comparing latent and active TB.

Kaforou *et al.* [135] conducted a microarray study to investigate the progression from latent to active TB in different disease backgrounds. The dataset comprised of 536 adults from South Africa and Malawi diagnosed with latent TB (n = 167), active TB (n = 194) and other non-TB diseases (with similar clinical phenotypes; n = 175) and were either HIV positive (n = 273) or HIV negative (n = 263). The authors then employed an

elastic net variable selection algorithm (a technique combining both the lasso and ridge regression methods) to calculate a disease risk score for each patient. The risk score would provide the basis for determining each patient's risk for developing active TB. Their primary analysis showed that clustering of samples (patients) was based on disease state – active TB, latent TB or 'other diseases' – rather than geographical location (South Africa or Malawi) or HIV status. As such, HIV infected and uninfected patient cohorts from South Africa and Malawi were grouped together to identify transcript signatures in patients with differing HIV statuses that are applicable across geographic locations. This risk score was, therefore, an indicator of an underlying signature of TB disease progression, independent of HIV status or geographical location.

In the present study, we sought to analyse the same microarray dataset as Kaforou *et al.*, but to use alternative analytical methods (OPLS-DA and ML methods) in order to specifically probe the progression from latent to active TB in HIV positive patients.

Traditional statistical tests have been widely used for identifying DEGs that can be used as biomarkers. Typically, a gene is defined as differentially expressed if the observed difference in expression levels between two experimental conditions is statistically significant [37]. DEGs are typically identified as genes whose differential expression between groups is significant to a specific threshold (usually $p < 0.05$) and user-defined thresholds for expression fold change (FC) to discern up-and down-regulated genes. Traditional differential expression (DE) analysis using gene expression data provided by microarray technology can make downstream analysis challenging due to the high dimensionality of the datasets. In recent years, multivariate statistical analyses have been developed and applied to microarray datasets. Orthogonal Projections to Latent Structures-Discriminant Analysis (OPLS-DA), for example, is widely used to analyse gene expression data to distinguish two groups. OPLS-DA is a powerful and robust modelling tool with the main benefit in interpreting data compared to other multivariate models being its ability to separate predictive from non-predictive (orthogonal) variation [136]. An additional advantage is the ability to determine the optimal number of orthogonal components, thus improving its cross-validated accuracy.

ML classifiers can be trained on gene expression data to find significant features and combining multiple ML methods to identify differentially expressed genes as biomarkers is an effective approach [137]. This combined approach takes advantage

of the strengths and weaknesses of the individual methods to improve biomarker discovery [137]. Reducing the number of features in machine learning approaches plays a crucial role when working with large gene expression datasets, including avoiding overfitting, speeding up training, and resulting in better classification through the reduced noise in the data [138]. When testing a large number of hypotheses, as is the case for high-dimensional data, where the number of features is larger than the available samples, a high number of false-positive test results can be expected [139]. Consequently, the raw *p*-values in the transcriptomic dataset need to be adjusted to control a false positive rate using the false discovery rate (FDR). The FDR measures the expected proportion of false positives among a set of hypothesis tests called significant [140]. Furthermore, the ML algorithms tend to perform better if the dataset is trimmed down. For this reason, a pre-filtering step is required to compare how filtering the dataset would affect the efficiency of the models.

The commonly used supervised machine learning algorithms that have previously been applied to transcriptomic datasets for disease studies include logistic regression (LR), support vector machine (SVM), decision tree (DT), random forest (RF) and artificial neural network (ANN). LR is the simplest to implement and mathematically the least complex. However, its accuracy declines when input variables have complex relationships. It is also vulnerable to overfitting and may overstate the prediction accuracy as a result of sampling bias. SVC is more robust than LR as it can handle multiple feature spaces, and it also has less risk of overfitting. However, it has the disadvantage of being computationally expensive for large datasets and may not perform well in the case of noisy data. The RF classifier consists of a set of decision trees derived from a randomly selected subset of the training set. The votes from different decision trees are aggregated to determine the output predictions [141]. RF can scale well for large datasets and can handle thousands of input variables. It also can provide estimates of which attributes are the most important for classification. It is one of the more complex ML algorithms, and as with SVM, it is computationally expensive. The RF algorithm is also easily prone to overfitting [142].

Since microarray datasets have a large volume and high dimensional data, it is normally expected that SVCs, RFs and neural networks outperform other classification algorithms, such as LR [143]. Both LR and SVC have been used in existing research for TB prognosis [143]. RF is currently one of the most used ML algorithms and has been widely applied in TB detection [142, 144].

Supervised ML algorithms can be combined with feature selection methods to identify marker genes from transcriptomic profiles. Feature selection is a primary step in machine learning as it helps improve the performance of the model being trained. Several feature selection methods are available that can be applied to different algorithms. Due to the size of a microarray dataset, the choice of feature selection methods is pragmatic, as some feature selection methods tend to be computationally expensive and time-consuming. In LR and SVC, standardised regression coefficients can be compared to determine feature importance. For most implementations of RF, the default feature importance choice is the mean decrease in impurity (MDI). The MDI of a feature is computed as the total reduction in loss or impurity contributed by all splits (across all decision trees) [145].

Using an integrative data-driven approach that leverages statistical DE results and results obtained by ML feature selection and classification can provide a viable gene signature that helps understand differences between disease states. Van Ijzendoorn *et al*. [146] combined statistical tests with machine learning techniques to reveal biomarkers and targets for soft tissue sarcomas. To select the genes of interest, they combined significantly differentially expressed genes (with a Benjamini-Hochberg adjusted *p*-value < 0.05 and logFC > 0) with RF. In addition, Abbas and El-Manzalawy presented a novel approach that leveraged i) statistical DE analysis to identify a list of DEGs (using absolute fold change ≥1.5 and adjusted p-value ≤0.05) ii) supervised feature selections methods (Random Forest Feature importance and minimum Redundancy and Maximum Relevance) to obtain an optimal subset of DEGs iii) supervised machine learning methods to evaluate the discriminatory power of the selected genes [147]. Their approach resulted in a 10-gene signature of mortality in paediatric sepsis. Both approaches mentioned above utilised supervised ML feature selection to refine the outcome of the statistical DE analysis. In this study, we applied statistical DE and supervised ML feature selection separately to an entire transcriptomic dataset and then integrated the outcome of the two methods to define a gene signature.

In the context of TB/HIV coinfection, Duffy *et al*. [148] used four previously published TB and HIV coinfection datasets to train and validate six machine learning classifiers (namely Random Forest, Support Vector Machine, Neural Networks, Elastic-net Logistic Regression, K-Nearest Neighbour and Extreme Gradient Boosting) to predict TB and HIV status. Their work generated a multinomial model that discriminates TB from non-TB states (including LTB and OD) and discriminates HIV positive TB as a

unique disease state. More recently, meta-analysis has been used to integrate transcriptome datasets from various studies to screen for TB biomarkers in HIV positive individuals [149]. Since these transcriptome-based biomarker studies focused solely on TB disease and HIV coinfection, the identification of predictive biomarkers for the progression of latent TB infection to active TB disease in HIV infected individuals remains unclear.

Furthermore, an ideal biomarker should be capable of discriminating between latent TB and active TB and be functionally relevant. For this reason, further enrichment analysis of DEGs is required to infer networks and associated pathways from expression profiles. An enrichment analysis can provide avenues for further investigation into the potential biological mechanisms of sets of genes.

Against this background, we present a novel approach that leverages statistical differential expression analyses (OPLS-DA) and supervised ML and feature selection methods to an entire transcriptomic dataset and integrate the outcome of the two pipelines to characterise progression from latent to active TB in HIV infected patients. The dataset used for this purpose is the microarray dataset of Kaforou *et al*. We identified a minimal gene signature, but also performed pathway enrichment analyses on a broader list of genes of interest to ascertain their biological function in relation to latent TB and HIV coinfection.

2.1.1 Aim

To develop a computational pipeline using statistical modelling and machine learning methods to identify genes of interest and a minimal gene signature associated with latent TB in HIV positive patients and to ascertain the biological functions associated with these genes.

2.1.2 Objectives

1. Apply OPLS-DA statistical modelling on the Kaforou *et al*. microarray dataset to find a list of differentially expressed genes that distinguish latent and active TB classifications.
2. Utilise machine learning and feature selection methods on the same dataset to identify differentially expressed genes distinguishing the latent and active TB classifications.

3. Integrate the outcomes of the statistical and machine learning approaches to define a minimal latent TB gene signature in HIV infected persons.
4. Perform a network enrichment analysis to identify biological pathways related to the genes of interest using the outputs from the statistical and machine learning methods.

2.2 Methods

2.2.1 Dataset

The microarray dataset of Kaforou *et al*. (NCBI GEO repository accession number GSE37250) was used for all analyses. The sample consisted of 536 adults from South Africa and Malawi diagnosed with latent TB (n = 167), active TB (n = 194) and other non-TB diseases (with similar clinical phenotypes; n = 175) and were either HIV positive (n = 273) or HIV negative (n = 263). The latter groups were combined to create the 'all patients' group. The dataset contained quantile-normalised gene expression data. Ilumina probe IDs were converted to Entrez gene IDs using the RStudio (v1.3.1093) BiocManager package. The expression changes between the latent and active TB groups were represented as $\log_2$ fold change (FC) in the HIV positive or 'all patients' groups, where FC = (counts for latent TB)/(counts for active TB).

2.2.2 Data Pre-Filtering

The SelectFdr class of sklearn (v0.24.1) in Python (v3.8) was used to filter the dataset using ANOVA with Benjamini-Hochberg (BH) correction for multiple comparisons. The data was filtered to either $p < 0.05$, $p < 0.01$ or $p < 0.001$. Probes associated with higher p-values were excluded in each case. FDR- and unfiltered datasets were used as input for ML classifiers.

2.2.3 Differential Expression Analysis using Orthogonal Projections to Latent Structures – Discriminant Analysis (OPLS-DA)

Prior to performing OPLS-DA, a principal component analysis (PCA) was done for data visualisation using the web-based tool MetaboAnalyst. OPLS-DA was conducted using the base packages pyopls (v0.24.1) and sklearn (v0.24.1) in Python (v3.8). The number of components for the OPLS-DA model was optimised for the HIV positive group and the 'all patients' groups, and a PLS scores plot and ROC curve was generated for the respective groups. A 34-component OPLS was performed to improve cross-validated accuracy from 0.8014 to 1 in the HIV positive patient group, and a 20-

component OPLS improved accuracy from 0.8604 to 1 in the HIV all patient group. Two sets of the top 100 DEGs ranked by PLS regression coefficients were obtained amongst the latent TB and active TB classes for both the HIV positive and 'all patients' groups.

2.2.4 Identification of Differentially Expressed Genes using a Machine Learning and Feature Selection Approach

ML algorithms (RF, LR or SVC) were trained on quantile-normalised expression counts (GSE37250) to which no filtering, or FDR-filtering (Section 1.2.2) had been applied. The full dataset was used to represent the 'all patients' group whereas the HIV positive cohort was used for the 'HIV positive' group. LR was regularised with an L2-penalty term. SVC and RF algorithms were optimised using 5-fold cross-validation on each dataset. A grid search across the hyperparameter space was employed for SVC ('C': [0.01,0.05,0.1], 'gamma': [5,4,3]) and a linear kernel was used. For RF, a random grid search was employed to optimise parameters ('n_estimators', 'max_features', 'max_depth', 'min_samples_split', 'min_samples_leaf', and 'bootstrap'). Data were standardized (Z-scaled) and classes ('latent TB' and 'active TB') encoded using the LabelEncoder function of sklearn. The data was split into training and testing sets in a 9:1 ratio and optimised ML algorithms were trained on the training set. AUC values relating to predictions made by models on the test set were captured as accuracy metric. For LC and SVC, model coefficients were standardised by multiplying each variable coefficient with the standard deviation of its expression value, and captured to reflect feature importance. Mean decrease in impurity (MDI) was used as a metric for RF feature importance and was calculated using the feature_importance function in sklearn. The training-testing-feature importance workflow was repeated 5000 times. After every 50 iterations, the top 10 genes by feature importance were captured. Notebooks detailing the procedures described above are available as supplementary material . The cumulative frequency for each gene was determined by adding the number of times it occurred in the Top 10 genes over the 5000 iterations.

2.2.5 Pathway Enrichment Analysis

A signalling network was constructed using the NetworkAnalyst visual analytics platform from the list of DEGs identified from the OPLS-DA analysis and the ML classifiers that were unique to the HIV positive group. SIGNOR 2.0 data were used to construct the network. The network was subjected to enrichment analysis for associations with terms in the Gene Ontology (GO), Kyoto Encyclopedia of Genes and

Genomes (KEGG) and Reactome databases. GO annotations were split into GO Biological Pathway (GOBP), GO Molecular Function (GOMF) and GO Motif (GOMo) terms. Terms that were significantly enriched (FDR $p < 0.05$) were considered. Significant terms that were shared by genes identified by OPLS-DA and all three ML classifiers were further considered and ranked according to average FDR *p*-value.

2.2.6 Integrating Machine Learning and Statistics-Based Approaches to Select Genes of Interest

A minimal gene signature was established by integrating the cumulative frequency Z-scores, the standardised OPLS-DA regression coefficient and log_2FC expression values of the three different ML classifiers. Genes of interest were selected if the cumulative frequency Z-score obtained from the ML and feature selection and the standardised OPLS-DA regression coefficients were in the upper quartile of the data and if the log_2FC expression values were either in the upper quartile (for upregulated genes) or the lower quartile (for down-regulated genes). The genes of interest from the three ML classifiers were consolidated to define two minimal gene signatures, one for the HIV positive cohort and the other for the 'all patients' cohort.

2.3 Results

To identify a transcriptomic minimal gene signature to distinguish latent and active TB classifications in HIV infected patients, OPLS-DA and ML modelling was performed on a previously published microarray dataset consisting of 536 patients in two geographically distinct African adult populations that have been diagnosed with either latent TB, active TB or other diseases and had either an HIV positive or HIV negative status [135]. Primary analysis conducted in the original study showed clustering by disease state (active vs latent TB) independent of HIV status and geographical location. Consequently, the 'experimental' group in the study reported here consisted of the entire HIV positive cohort from both African populations and the 'control' group, referred to as the 'all patients' group used in this study was the total cohort, irrespective of geographical location or HIV status.

2.3.1 Data Exploration using Principal Component Analysis (PCA)

A standard two-component PCA model was applied to explore the natural separation patterns seen in the microarray data. The first and second principal components (PC1 and PC2) of the HIV positive group explained only 5.3% and 3.6% of the variation,

respectively (see Figure 2.1). Similarly, PC1 and PC2 in the 'all patients' group explained only 4.9% and 3% of the data, respectively. Although a modest separation was observed between the two classes (active TB and latent TB), the two-component PCA was not sufficient in explaining the variation in the data, and therefore could not model transcriptomic differences between latent and active TB in the different groups; as such, more refined approaches were explored to interrogate the data.

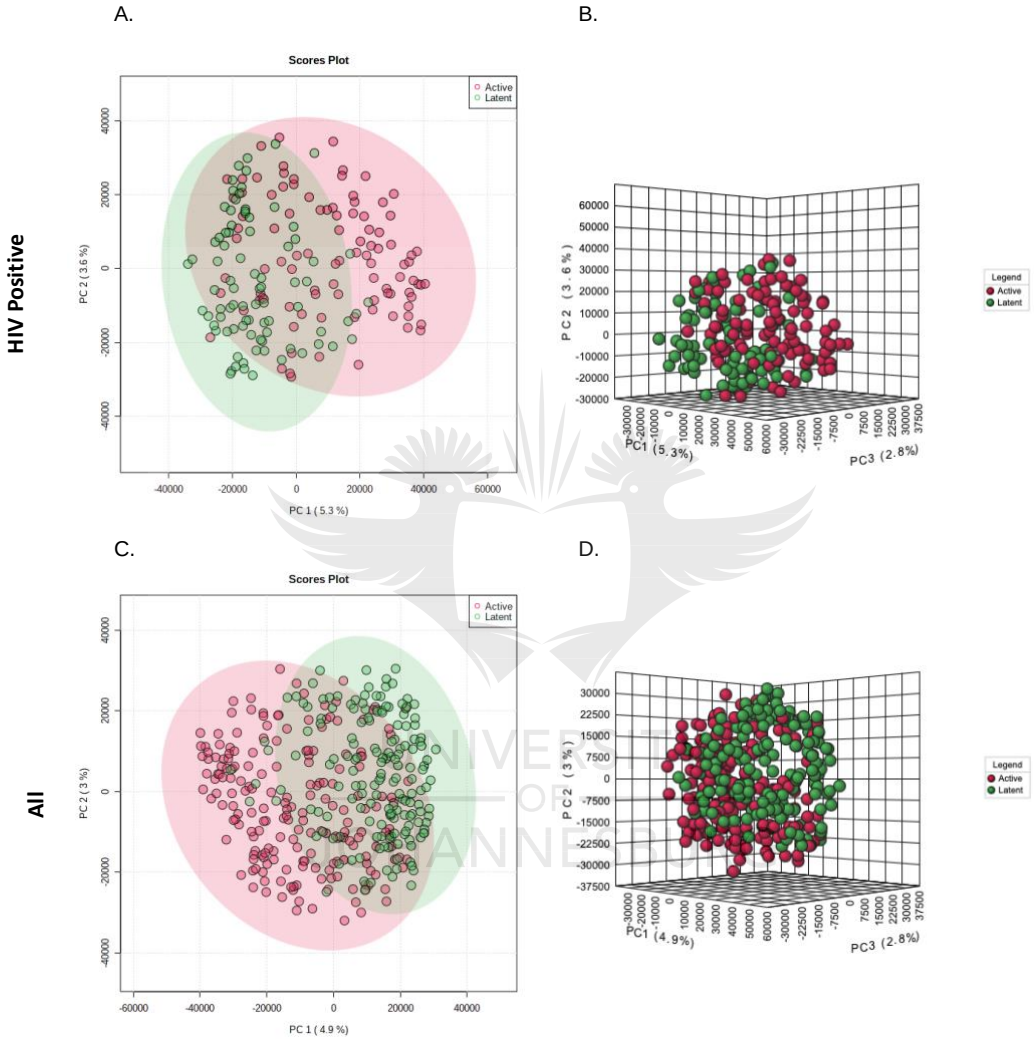


Figure 2.1 Explorative analysis of microarray dataset using a two-component PCA model.
(A) PCA score plot of the two classes (active TB vs LTB) in HIV positive patients where PC1 and PC2 explained 5.3% and 3.6% of the variation, respectively. (B) 3D score plot of the two classes (active TB vs latent TB) in HIV positive patients. (C) PCA score plot of the two classes (active TB vs latent TB) in all patients where PC1 and PC2 explained 4.9% and 3% of the variation. (D) 3D score plot of the two classes (active vs latent TB) in all patients. Figures were generated in Metaboanalyst.

2.3.2 Differential Expression (DE) Analysis using Orthogonal Projections to Latent Structures Discriminant Analysis (OPLS-DA)

To more precisely investigate the structure of the data and identify differentially expressed genes, OPLS-DA modelling was employed. This modelling process

involved a two-step procedure. The first step entailed optimising the number of orthogonal components, and PLS regression analysis was conducted in the second step. The score plots showed an obvious separation of the active and latent TB classes in both HIV positive and 'all patients' groups (Figure 2.2 A and C). The model for HIV positive patients required 34 components, while a 20-component model was used for the 'all patients' group. These models accurately explained the data variance, as is evident from the ROC curves with AUC = 1 (Figure 2.2 B and D). This high AUC score indicated the models' ability to distinguish between the active TB and latent TB classes perfectly. The OPLS-DA model yielded 1648 DEGs for the HIV positive group and 2167 DEGs for the 'all patients' group from a total of 5574 genes. The genes of interest were ranked according to their PLS regression coefficients (the top 200 DEGs for each group are shown in Table S1 in the Supplementary material).

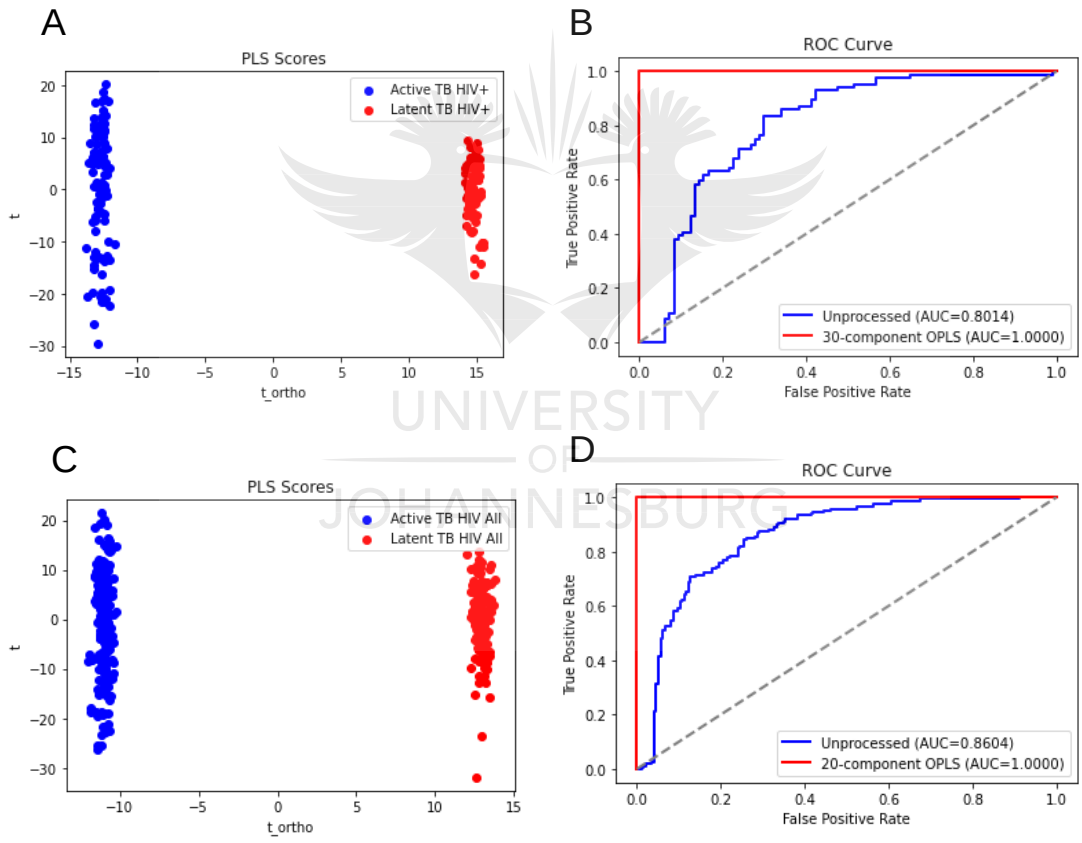


Figure 2.2 OPLS-DA score plots and ROC curves showing the discriminant separation between active and latent TB classes in HIV positive and 'all patients' groups.
OPLS-DA score plots of the two classes (active TB vs latent TB) in HIV positive patients (A) and the 'all patients' group (B) are shown. ROC curves of the two classes (active TB vs latent TB) in HIV positive patients and the 'all patients' group are shown in (C) and (D), respectively.

2.3.3 Identification of DEGs using a Machine Learning and Feature Selection Approach

As an alternative means to identify DEGs that play a role in the progression of disease from latent to active TB in HIV positive patients, machine learning (ML) algorithms were investigated. We developed an ML workflow to identify genes predictive of active or latent TB (Figure 2.3). Using pre-filtered expression data as an input, each machine learning algorithm generated an output of either active TB or latent TB classifications based on the expression pattern of the microarray dataset. The raw *p*-values in the transcriptomic dataset were adjusted to control a false positive rate using the false discovery rate (FDR). Three FDR *p*-values (<0.05, <0.01 and < 0.001) were applied to the dataset as pre-filtering steps to improve the performance of the ML algorithms. These FDR-filtered datasets were compared against the unfiltered dataset in the three ML analyses. Supervised ML algorithms, namely RF, LR and SVC, were applied to the training dataset to classify samples as either latent or active TB given the pre-filtered or unfiltered microarray expression data. Each machine learning algorithm was optimised by cross-validation. Following optimisation, the Top 10 discriminative features were extracted. For LR and SVC, the regression coefficients were used to rank differentially expressed genes (i.e. as a feature selection method), while MDI attributes were used for this purpose when RF was applied. This procedure was repeated over 5000 iterations. The cumulative occurrence (frequency) of genes in the top 10 discriminating features of each of the 5000 ML models generated were used as a final gene ranking metric.

To compare the accuracy of the three machine learning classifiers applied to the differentially filtered input datasets, the performance of the classifiers was assessed using the area under the ROC curve (AUC) as the performance metric (Figure 2.4).
The pre-filtering of data using an FDR *p*-value < 0.05 yielded the narrowest distributions and highest median AUC scores: 0.9296, 0.9264 and 0.926 for the LR, RF and SVC models, respectively, in the 'all patients' group. In comparison, the reported AUC values for the HIV positive group included 0.9169, 0.9017 and 0.9163 for the LR, RF and SVC models, respectively. An ANOVA test confirmed that the FDR p <0.05 pre-filtration yielded statistically significant results. Subsequently, the pre-filtering method of FDR < 0.05 was used for the rest of the analyses reported in this Chapter.

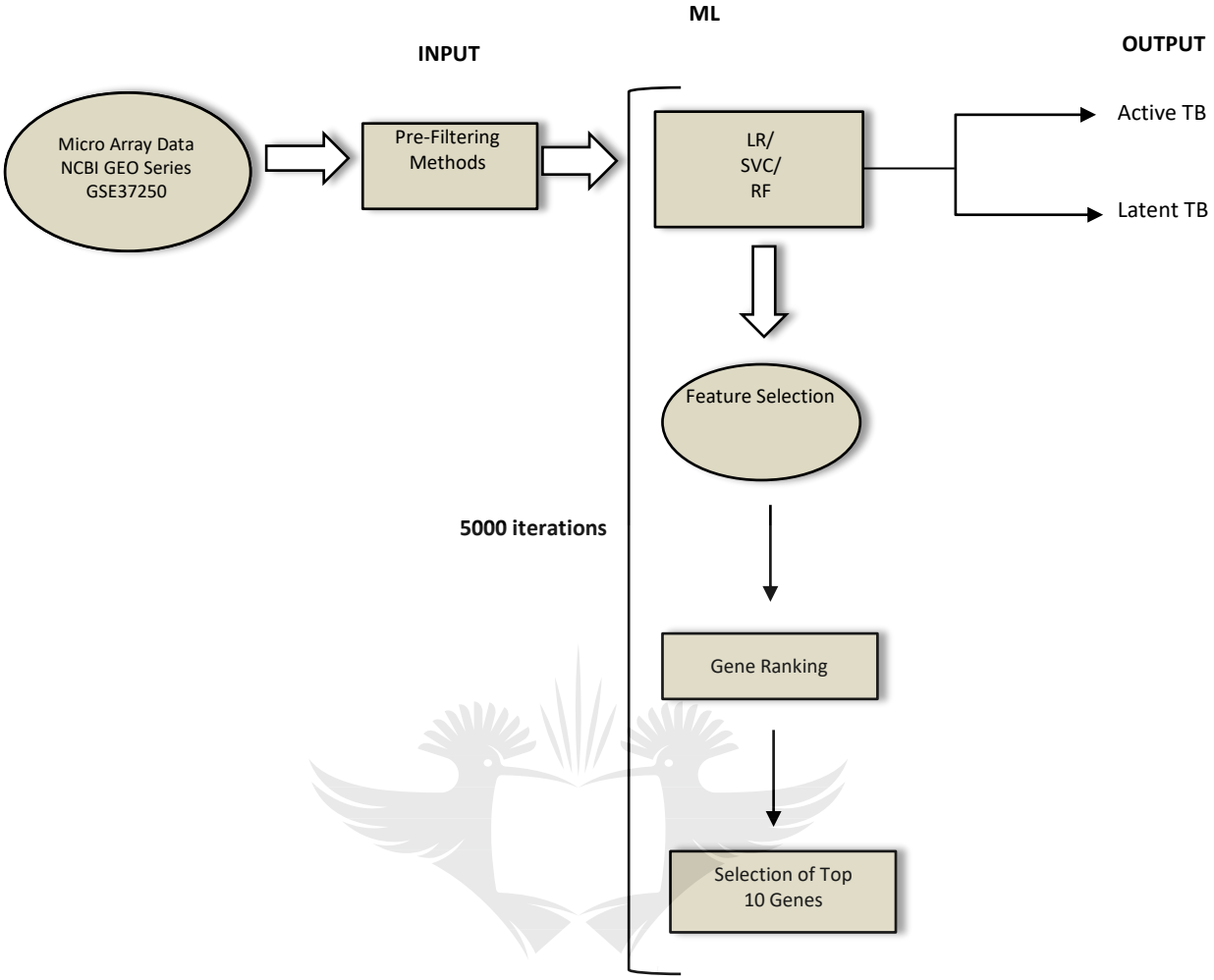


Figure 2.3 Machine learning pipeline for the selection of DEGs. The Logistic Regression (LR), Support Vector Classifier (SVC) and Random Forest (RF) machine learning algorithms were applied to the pre-filtered microarray dataset to classify latent TB and active TB DEGs. Following training, the Top 10 features were extracted, and the genes were ranked accordingly.

The median AUC of the models ranged from 90.2% to 91.7% in the HIV positive group and 92.4% to 93.0% in the 'all patients' group. Although the algorithms exhibited high performance as assessed by their comparable AUC values, LR marginally outperformed the other two algorithms, while RF showed a slightly lower performance over the 5000 iterations. However, this is not statistically significant as judged by an ANOVA test ($p > 0.05$).

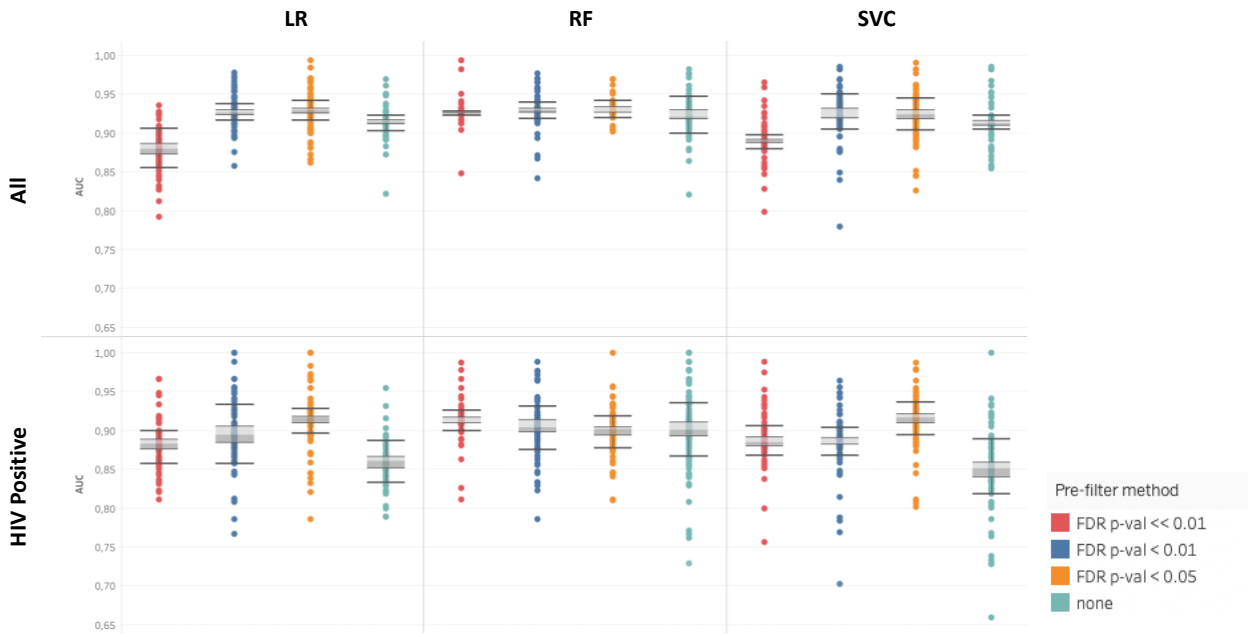


Figure 2.4 Comparison of the performance of the different ML classifiers in the two patient groups. Each box and whisker plot represents the distribution of the average AUC score of the ML classifiers across three different levels of FDR pre-filtering. The box consists of the upper quartile, lower quartile and the median in the centre, along with the upper and lower hinge. The range between the upper and lower quartile values are shown as the dark shaded regions. The whiskers indicate variability outside the upper and lower quartiles.

After each iteration, the top 10 genes of which the expression levels could discriminate between active and latent TB were determined by feature importance metrics and ranking. The cumulative number of times that a gene appeared in the top 10 across the 5000 iterations was determined and used to rank the DEGs identified in this manner (Figure 2.5). Genes with the highest cumulative occurrence (frequencies) in the top 10 were deemed as the most significant contributors to a transcriptomic signature of latent vs active TB. Within the HIV positive cohort, there were a total of 121, 124 and 116 genes occurring in the top 10 using the LR, SVC or RF classifiers, respectively. While the 'all patients' group resulted in 118, 145 and 55 genes occurring in the top 10 from the LR, SVC or RF classifiers, respectively.

Although each algorithm yielded several unique genes, there was an overlap in the genes found by all three classifiers (Figure 2.6, Table 2.1). The relatively small extent of overlap may reflect the distinct mathematical underpinnings of each classifier. Still, it may equally express the intertwined and redundant nature of the biological pathways that give rise to latent vs active TB phenotypes.

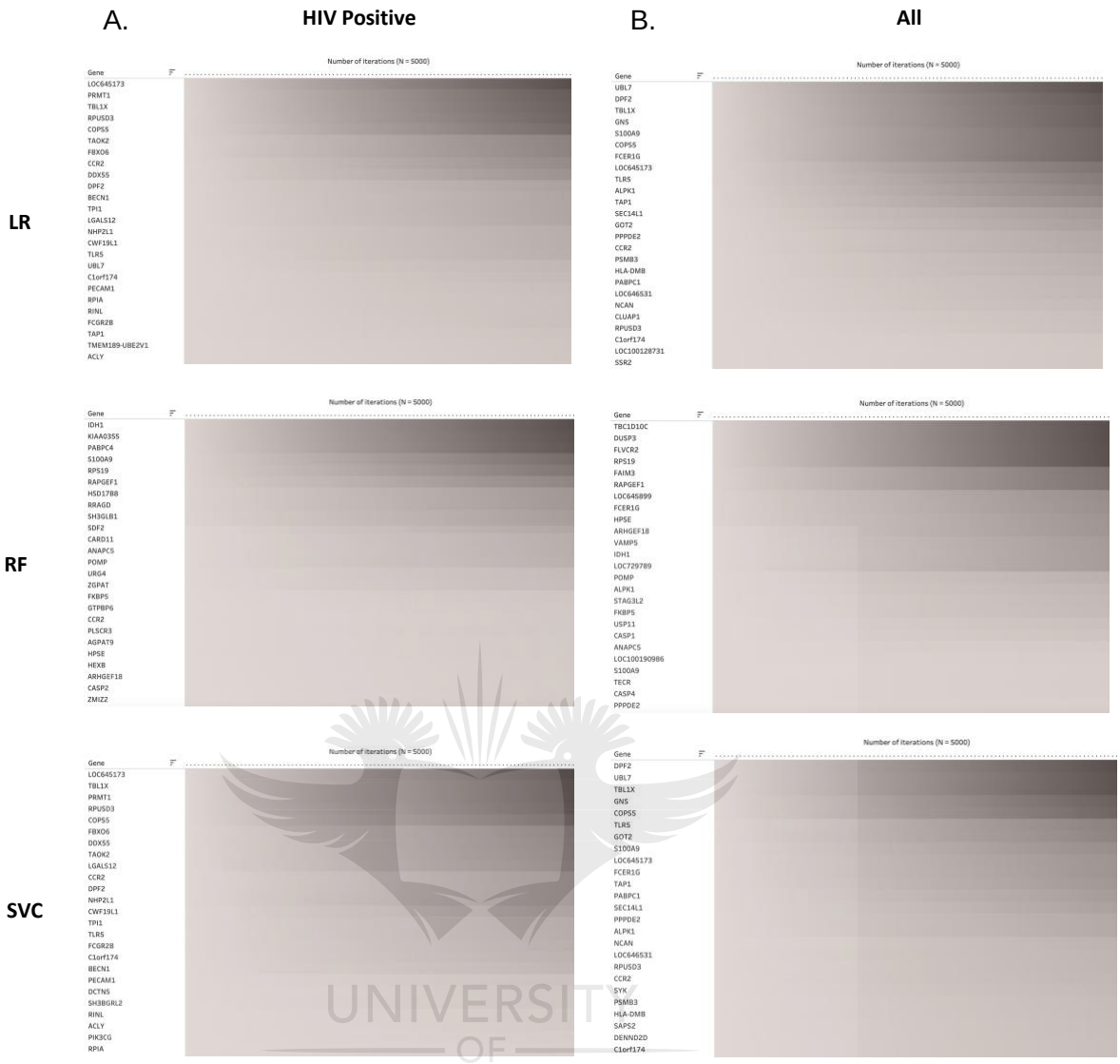


Figure 2.5 Cumulative frequencies of the genes that appear in the top 10-ranked genes across 5000 iterations of the feature selection process following ML modelling with pre-filtering at FDR p-value < 0.05.

Classifiers acting on the HIV positive group are shown in (A) and those acting on the 'all patients' group are shown in (B). All figures show only the top 25 genes. The cumulative frequency of occurrence in the top 10-ranking genes is colour-coded. The darker the intensity of grey, the more times the genes cumulatively appeared in the top 10-ranking genes.

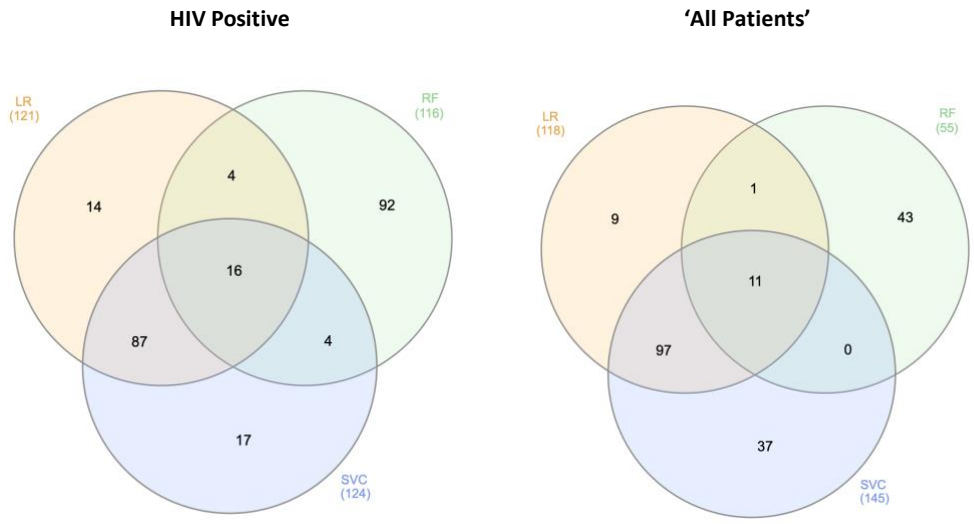


Figure 2.6 Relationship between sets of top-ranking genes identified by Logistic Regression (LR), Random Forest (RF) or Support Vector Classifier (SVC) in the HIV positive and 'all patients' groups. Venn diagrams were generated using InteractiVenn.

Table 2.1 Genes symbols and cumulative frequencies of top genes occurring in the output of all three ML approaches. Genes in bold are unique to either the HIV positive or the 'all patients' cohort. The frequency reported is the average cumulative frequency of occurrence after application of the three ML classifiers.

HIV Positive		*All patients*	
Gene Name	*Cumulative Frequency*	*Gene Name*	*Cumulative Frequency*
RASSF5	9	**POMP**	1226
LOC644132	20	**GADD45B**	131
GOT2	27	**RAPGEF1**	2088
FLVCR2	111	**IDH1**	1060
ECH1	20	**TRPV2**	273
ALPK1	611	CCR2	987
LOC653314	64	PPPDE2	948
PPPDE2	329	ALPK1	1515
CARD11	894	LOC645173	2207
MMS19L	172	FCER1G	2507
FCER1G	962	**S100A9**	1939
TAP1	343		
DPF2	1792		
CCR2	1679		
RPUSD3	3727		
LOC645173	3646		

A Pearson correlation matrix was generated to investigate the correlation between the classification models and how well the algorithms performed with respect to one another. To identify which classifier performed best at detecting differences between the two patient groups, an upper cut-off of 0.3 for the Pearson correlation coefficient value (*r*) was applied since a value lower the 0.3 indicates a weak correlation. The correlation between HIV positive and 'all patients' groups was the lowest for the RF classifier (*r*-value of 0.1446), suggesting a weak correlation between the genes selected to classify active TB and latent TB. In contrast, the correlation was moderate between the groups for the LR (*r*-value of 0.3816) and SVC (*r*-value of 0.4236) classifiers. The corresponding *r* values imply that the RF classifier best differentiated between the HIV positive and 'all patient' groups.

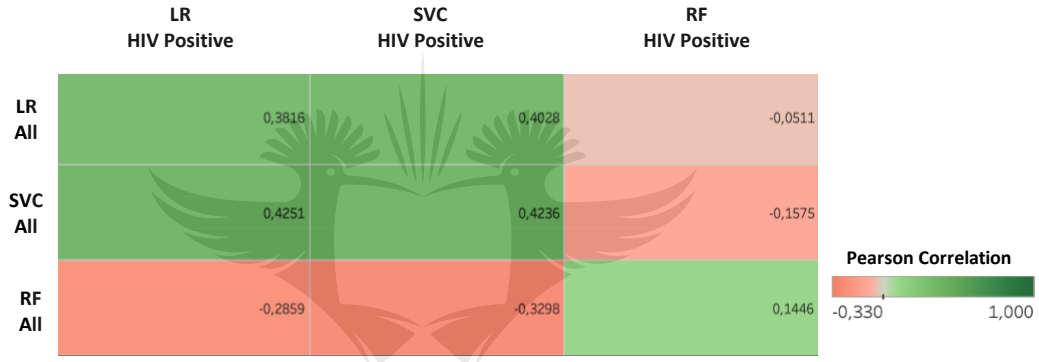


Figure 2.7 Pearson correlation matrix of ML classifiers in the different groups using a pre-filtration at FDR p-value < 0.05.

2.3.4 Pathway Enrichment Analysis

Pathway enrichment analysis is a network-based method that provides insight into gene lists generated from omics data by finding functional categories such as overly represented pathways in an experimental set. Utilising this method allowed an exploration of the functional context of the genes of interest obtained using OPLSDA and the different ML classifiers. The main objective of this analysis was to understand if the function of the genes obtained was relevant to latent TB infection to corroborate our approach to obtain genes relevant to the progression of latent TB to active TB.

Genes obtained from the OPLS-DA and the ML classifiers unique to the HIV positive group were used to construct a signalling network using NetworkAnalyst (Figure 2.8). These DEGs were then subjected to enrichment analysis using five databases, namely GOBP, GOMF, GOMo, KEGG and Reactome. Figure 2.9 presents the unique and overlapping pathway enrichment terms distinguished from the OPLS-DA and the

machine learning models across the five databases. Overlapping terms were ranked by statistical significance based on the average FDR *p*-value.

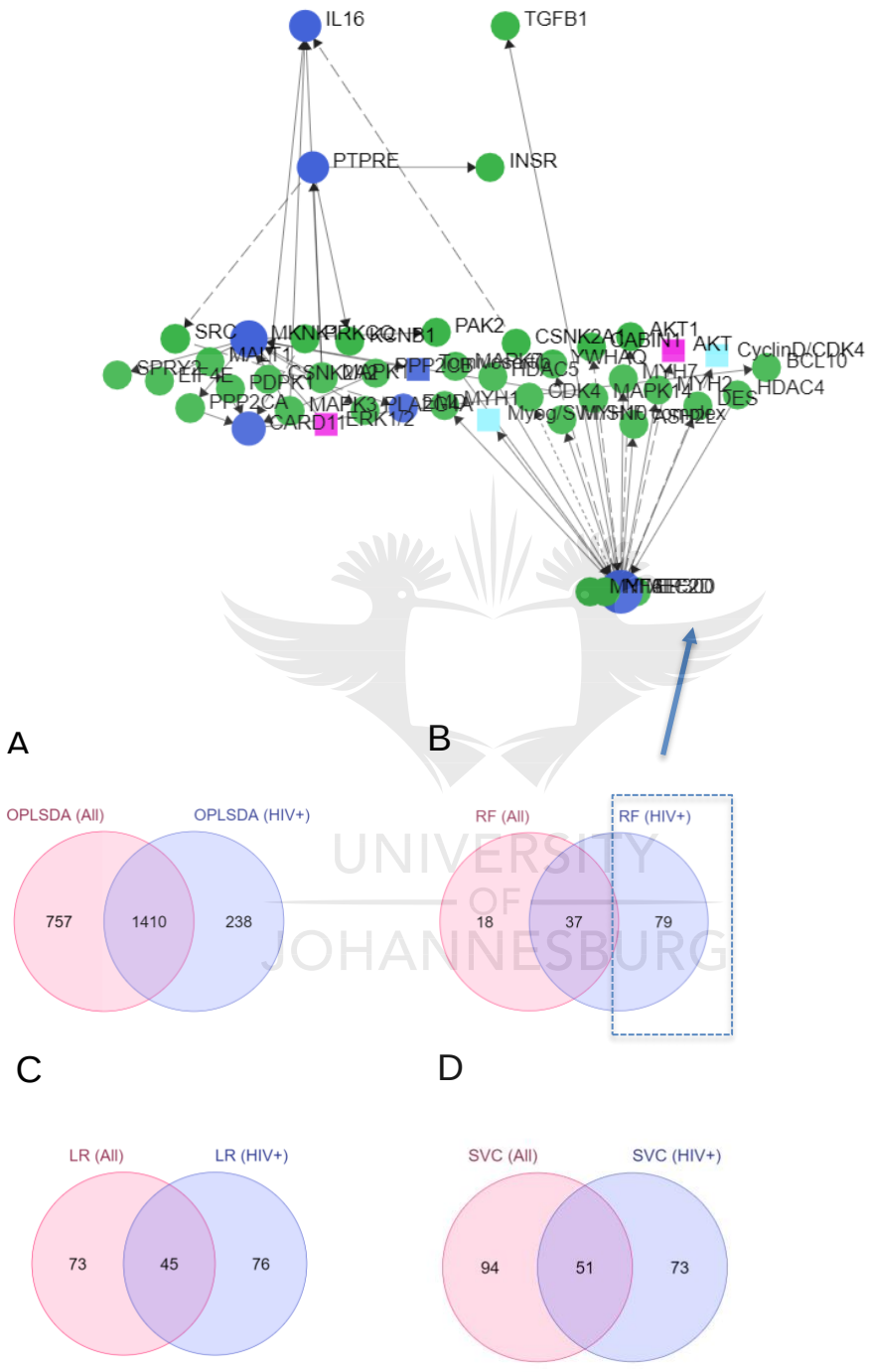


Figure 2.8 Example signalling network constructed using the gene set of the HIV positive group yielded by the RF model.
Venn diagrams and minimal triangular matrices represent the overlap of genes in the HIV positive and 'all patients' groups identified using the (A) OPLS-DA model (B) RF classifier (C) LR classifier, or (D) SVC classifier. Only the genes exclusive to HIV positive groups in the different classifiers were selected and used for further analyses. The signalling networks constructed using the gene lists from the different ML classifiers can be found in the supplementary section (Supplementary Figure S1-S4). The signalling

networks were constructed using **NetworkAnalyst**. Venn diagrams and minimal triangular matrices were constructed using Molbiotools.

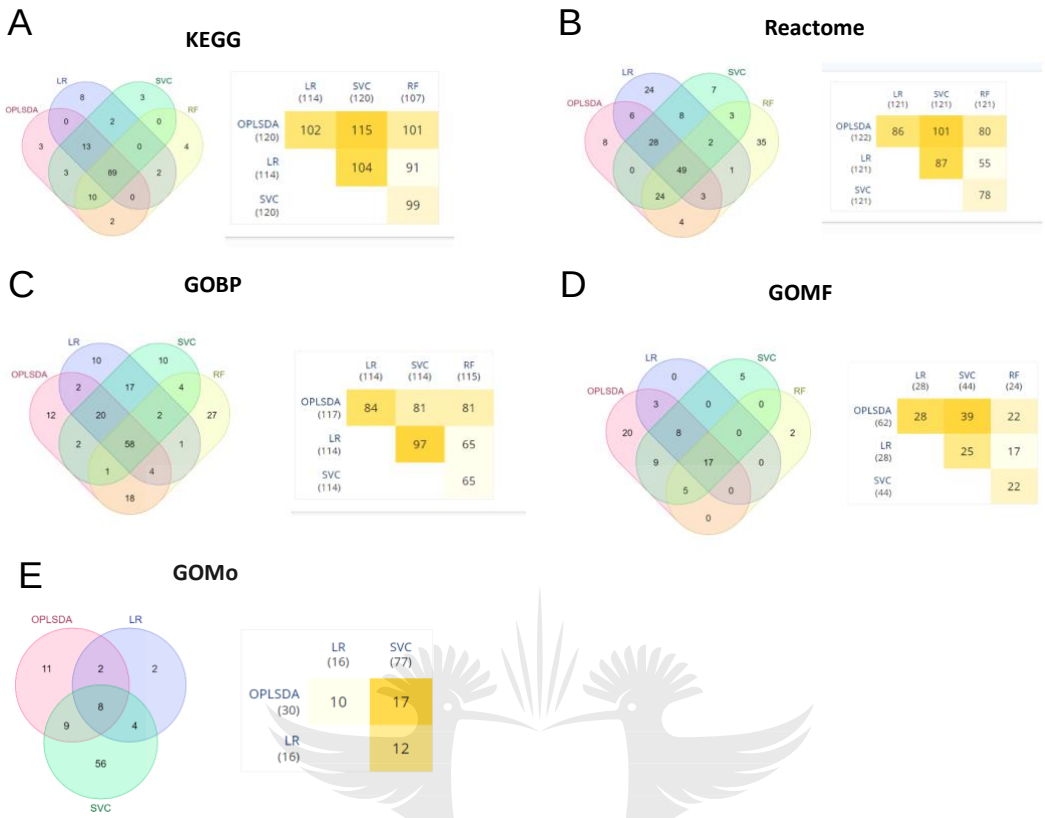


Figure 2.9 Comparison of the combined enriched pathway terms identified from the KEGG, Gene Ontology, and Reactome databases.
The Venn diagrams represent the unique and overlapping pathway enrichment terms determined from the OPLS-DA and ML models across the (A) KEGG pathway (B) Reactome pathway (C) GOBP pathway D) GOMF pathway (E) GOMo pathway. The triangular matrices constructed using Molbiotools represent the number of overlapping terms by pairwise comparison.

The gene set enrichment analysis identified specific pathways associated with latent TB to active TB progression in the HIV infected cohort. The top pathways found in the enrichment analysis that have been reported in literature within the context of active TB, latent TB or TB/HIV coinfection are summarised in Table 2.2. The detailed list of the comparison of the pathway enrichment terms found using the GO, KEGG and Reactome databases can be found in Supplementary Table S4.

Table 2.2 The top enriched pathway terms that have been reported in the context of latent or active TB in literature were determined from the OPLS-DA and ML models in HIV positive patients using the combined KEGG and gene ontology databases.

The represented pathways are ranked according to statistical significance. Only pathway terms with an FDR $p < 0.05$ were considered statistically significant.

Top Pathways	*OPLSDA*	*LR*	*SVC*	*RF*	*Avg FDR*	*Literature report in the context of latent TB/active TB or TB/HIV coinfection*
			FDR p-Value			
GGGCGGR_V$SP1_Q6	1.08E-02	1.78E-02	6.61E-03	-	1.17E-02	The GGGCGGR motif matches the annotation for the Sp1 transcription factor [150]. Literature has shown that the recruitment of Sp1 to the TNF-α promoter correlates to the critical functional roles that the Sp1 promoter site plays in the activation of the gene by *Mtb* [151].
Signal Transduction	1.33E-05	1.76E-03	1.41E-06	3.25E-02	9.09E-04	The expression of cell signalling transduction receptors (such as CD14, TLR2, CD206, and β2 integrin LFA-1) have been studied on monocytes from patients with active TB and healthy individuals with *Mtb* latency. A simultaneous increase in the expression of the mCD14 receptor and LFA-1 integrin in active TB patients might be considered an early sign of breaking immune control by *Mtb* bacilli in subjects with latent TB [152].
Insulin receptor signalling cascade	7.87E-08	4.77E-05	7.33E-08	2.00E-02	7.24E-04	Vitamin D is known to play a crucial role in the control of TB infection. Through the modulation of insulin resistance and secretion, vitamin D is related to controlling blood glucose in Type 2 Diabetes (T2D). Vitamin D deficiency is common in T2D patients, and some studies have related this deficiency to an increased risk of TB. Insulin signalling involves reduced antigen-specific proliferation and proinflammatory cytokine production in T cells. Diminished insulin production and modified insulin receptor-mediated signalling in T cells could increase TB risk in patients with diabetes of longer duration [153].
IRS-related events	7.87E-08	4.77E-05	7.33E-08	2.00E-02	5.96E-04	TB associated immune restoration syndrome (IRS) is a frequent event observed in approximately 10-30% of patients coinfected with HIV-1 and *Mtb*. TB-IRS is associated with an increase in the number of IFN-γ producing tuberculin specific cells[154].
ATP binding	1.12E-14	5.46E-07	3.54E-24	1.12E-14	3.79E-04	Bedaquiline is a drug used to treat active TB, particularly in individuals with multi-drug resistant TB. Bedaquiline targets the mycobacterial ATP synthase, a crucial enzyme in the obligate aerobic *Mycobacterium* genus; however, how ATP synthase binds to the enzyme remains unknown [155].
Innate Immune System*	1.05E-07	5.49E-05	1.04E-07	2.14E-02	3.23E-04	Cell-mediated immunity is crucial from controlling *Mtb* infection. The activation of $CD4^+$ and $CD8^+$ T cells occurs in active TB. The depletion of $CD4^+$ cells is characteristic of HIV and contributes to the increased risk of reactivation of latent TB. In the latent phase of TB several immune mechanisms including increased levels of $FoxP3^+$ Treg cells, TGF-β, IL-27, SOCS1, PGE-2 or decreased levels of TNF, IFN-γ and polyfunctional; specific T cells are involved in latent TB reactivation [156].
CD28 co-stimulation	9.47E-19	7.66E-07	5.37E-13	3.70E-04	1.36E-04	A recent study with B7DKO mice that were highly susceptible to chronic mycobacterial infection highlighted the importance of the CD28/B7 co-stimulation pathway, this susceptibility being due to impaired Th1 T-cell responses [157].
Endocrine resistance	3.11E-13	5.73E-09	1.64E-16	1.89E-04	1.85E-05	Although rare, TB of endocrine glands, such as the adrenals, thyroid, and pituitary, have clinically significant pathophysiologic effects; however, diagnosing endocrine gland involvement in TB is a clinical challenge. It is now known that even in the absence of direct gland involvement, endocrine and metabolic derangement can occur due to the TB disease process and/or anti-TB treatment [158].
TGF-beta signalling pathway	3.80E-04	1.41E-02	3.80E-05	5.81E-03	1.50E-05	There is robust activation of pathways downstream of TGF-β signalling among CD4+ T cells of the granuloma. This, along with previous literature on the deleterious role of TGF beta in TB, suggest the potential role of TGF beta signalling in curbing T-cell function. In addition to TLR2 and cytokine signalling, research has indicated TGFβ mediated signalling responses to be highly active in latent TB. TGFβ levels are also known to be high in active TB and are required for intracellular survival of *Mtb* [5].
PI3K-Akt signalling pathway	1.88E-04	1.02E-02	7.56E-05	9.99E-05	1.37E-05	Earlier studies have shown that *Mtb* and its components can trigger the PI3K/Akt signalling pathway in macrophages. It is possible that *Mtb*-mediated Akt signalling plays a major role in the suppression of macroautophagy during infection. PI3K-Akt signalling pathway has a key role in cell growth, differentiation, apoptosis, autophagy, metabolism and infectious disease, particularly tuberculosis [159].
Insulin signalling pathway	6.47E-09	1.59E-05	8.15E-09	1.81E-03	1.27E-05	Studies in mice have found that TB to be associated with increased insulin signalling and systemic glucose tolerance in adipocytes. Infection with *Mtb* stimulated adipose tissue inflammation and adipocyte hypertrophy, which are typically linked with insulin resistance [160]..
Protein serine/threonine kinase activity*	1.08E-10	1.51E-03	8.19E-20	1.08E-10	1.19E-05	PknE, a Serine/Threonine Protein Kinase of *Mtb,* plays a vital role in MAPK cross talks, enabling intracellular survival of *Mtb,* a survival strategy found to also affect HIV/TB co-infection[161].
Phosphotransferase activity, alcohol group as acceptor	4.88E-10	5.07E-04	9.98E-09	4.88E-10	8.05E-06	Lcp1 has been identified as an essential phosphotransferase that ligates two essential cell wall macromolecules found in the mycobacterial cell wall, namely arabinogalactan and peptidoglycan. The discovery of this phosphotransferase sheds new light on the final stages of the mycobacterial cell wall assembly. It suggests a key biosynthetic step that could be used for new anti-TB drug discovery [162].
Sphingolipid signalling pathway	9.32E-09	1.82E-08	7.34E-20	1.99E-02	6.75E-06	Sphingosine-1-phosphate (S-1P) is a key sphingolipid involved in the pathobiology of various respiratory diseases. Studies have illustrated the importance of S-1P in controlling non-pathogenic mycobacterial infection in macrophages. There also appears to exist the therapeutic potential of S-1P against pathogenic *Mtb*.[163].

Signalling by FGFR in disease	9.34E-14	1.28E-08	3.55E-10	4.15E-02	5.75E-06	A recent study has indicated that in addition to TLR2 and cytokine signalling, TGFβ, PDGFR, FGFR and EGFR mediated signalling responses to be highly active in latent TB infection [5].
MAPK signalling pathway	3.10E-12	3.31E-03	1.45E-04	1.74E-04	4.25E-06	MAPK is a major immune signalling pathway that mediates the regulation of innate immune responses by controlling the synthesis of different cytokines such as MAPK and NF-κB. The MAPK signalling pathway can also regulate the production of many cytokines by the macrophages infected with *Mtb* [164].
Fc epsilon RI signalling pathway	8.09E-12	1.62E-07	6.70E-15	4.20E-04	4.23E-06	After JieHeWan (JHW is a traditional Chinese medicine that exhibits anti-TB effects) treatment, one of the pathways related to the immune and inflammatory response regulation was the down-regulated Fc epsilon RI signalling pathway [165].
VEGF signalling pathway	1.71E-10	1.82E-08	4.89E-15	3.91E-02	2.14E-06	VEGF is an immunosuppressive that can inhibit the function of T cells, increase the recruitment of regulatory T cells (Tregs) and myeloid-derived suppressor cells (MDSCs), and hinder the differentiation and activation of dendritic cells (DCs). Research has reported that TB granulomas also have a functionally abnormal vasculature with enhanced expression of vascular endothelial growth factor (VEGF) and that anti-VEGF therapies could play a role in treating TB [166] through creating a normalised granuloma vasculature by blocking VEGF signalling [167].
FoxO signalling pathway	8.09E-12	1.62E-07	6.70E-15	4.20E-04	1.34E-06	FOXO3 (a target of the PI3KAkt pathway) has been proposed as a potential target for developing host-directed strategies for better prevention of or treatment of TB [168].
B cell receptor signalling pathway	4.36E-10	1.13E-06	4.34E-11	1.69E-05	6.19E-07	B cells play a crucial role in regulating innate and adaptive immune responses to infectious agents even in disease states dominated by T lymphocytes, as is the case of TB. Studies have shown B cell-deficient mice and human patients receiving B cell depletion therapy typically present changes in their $CD4^+$ T cell and $CD8^+$ T cell repertoires [169]. B cells in TB disease have been demonstrated in mouse models whereby B cell-deficient mice are more susceptible to TB [170]. *Mtb* membrane antigens are strong inducers of B cell responses resulting in the production of high antibody titres [171].
Apoptotic process	5.97E-22	1.57E-07	2.68E-14	2.78E-08	3.23E-07	Many studies have supported the model that virulent *Mtb* inhibits apoptosis, while avirulent *Mtb* induces apoptosis [172].
Protein phosphorylation	6.28E-18	5.04E-08	5.58E-18	3.11E-05	2.65E-07	Protein phosphorylation plays a key role in the physiology and pathogenesis of *Mtb*. The PtpA (a secreted protein tyrosine phosphatase) is a substrate of the protein tyrosine kinase and has proven essential for *Mtb* inhibition of host macrophage acidification and maturation [173]
Positive regulation of cellular metabolic process	6.78E-23	3.06E-08	2.74E-16	4.25E-06	2.44E-07	Recent studies have provided a better understanding of the metabolic interplay between host immune cells and pathogens and how their interactions impact disease outcomes and antibiotic-treatment efficacy. The metabolic cascades in immune environments during *Mtb* infection and the metabolites produced exhibit critical roles in the induction of anti-*Mtb* protective immunity and progression of disease [174].
Positive regulation of transcription, DNA dependent	5.84E-08	8.28E-04	2.22E-05	5.84E-08	1.37E-07	Rip1 is a kinase that acts in multiple signalling pathways to regulate inflammatory responses and trigger apoptosis and necroptosis. Although the substrates of Rip1 are undefined, it has been reported as a determinant of *Mtb* cell envelope composition and virulence during HIV coinfection. The protein has proven to positively regulate transcription of other proteins, namely BCG2962c and BCG2953 [175].
Transmembrane receptor protein tyrosine kinase signalling pathway	4.51E-16	2.48E-08	1.61E-12	5.28E-04	4.62E-08	Individuals who are latently infected with TB and healthy controls have shown differential expression in genes belonging to the regulation of metabolism, translation, apoptosis and signal transduction pathways that involve MAP kinase phosphatase and protein tyrosine/threonine phosphatase activities [176].
T cell receptor signalling pathway*	8.26E-09	1.36E-02	2.35E-10	6.54E-03	5.72E-11	*Mtb* modulates host immune response, particularly in T cell responses, for its survival, leading to disease or latent infection [177]. It has been shown that the glycolipids of the *Mtb* can indirectly inhibit $CD4^+$ T cells by interceding with T cell receptor signalling [178]. HIV infection may result in the upregulation of inhibitory receptors on *Mtb* $CD4^+$ T cells. This mechanism has been linked with antigen-specific T cell dysfunction in chronic infections [179].

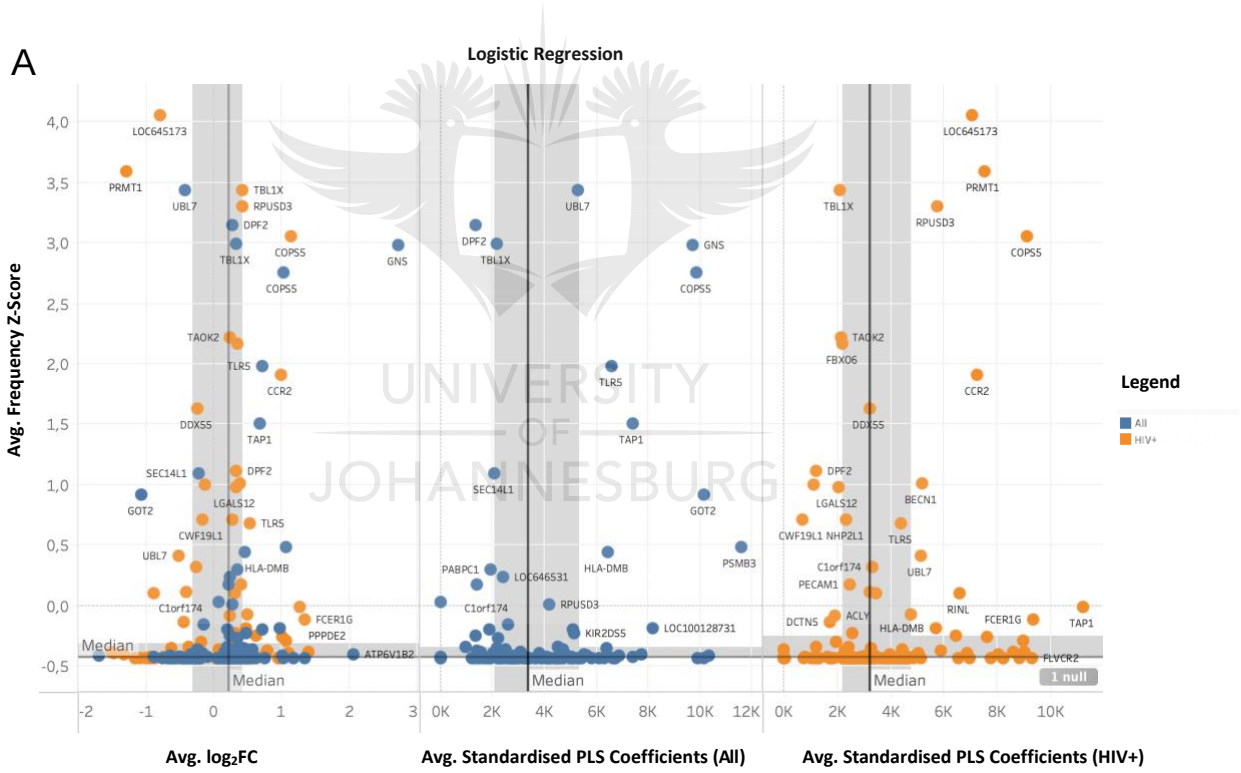
**Refer to pathways that have been linked to TB/HIV coinfection in literature studies.*

2.3.5 Integrating Machine Learning and Statistics-Based Approaches to Select Minimal Transcriptomic Signature

The results from the statistical and machine learning pipelines were integrated to identify a minimal transcriptomic gene signature relevant to the progression of latent to active TB (Figure 2.10). Z-scores of the cumulative frequencies (with which a gene was ranked in the top 10 following ML feature selection across 5000 iterations) were calculated in order to cross-compare data-sets. A gene was considered significant if :

1. The cumulative frequency Z-score was in the upper quartile of the data
2. The standardised OPLS-DA regression coefficient was in the upper quartile of the data
3. The log_2FC value was either in the upper quartile (for upregulated genes) or the lower quartile (for down-regulated genes)

Genes obtained from all three ML classifiers were combined, resulting in a total of 16 genes (eleven upregulated and five downregulated; Table 2.6) in the 'all patients' cohort and an 11-gene signature (seven upregulated and four downregulated; Table 2.5) for HIV positive individuals. A total of 5 significant genes were identified using the LR and SVC classifiers in the HIV positive group and four genes in the 'all patients' group. While six genes were discovered using RF in the HIV positive group, RF yielded only one gene of interest in the 'all patients' group.



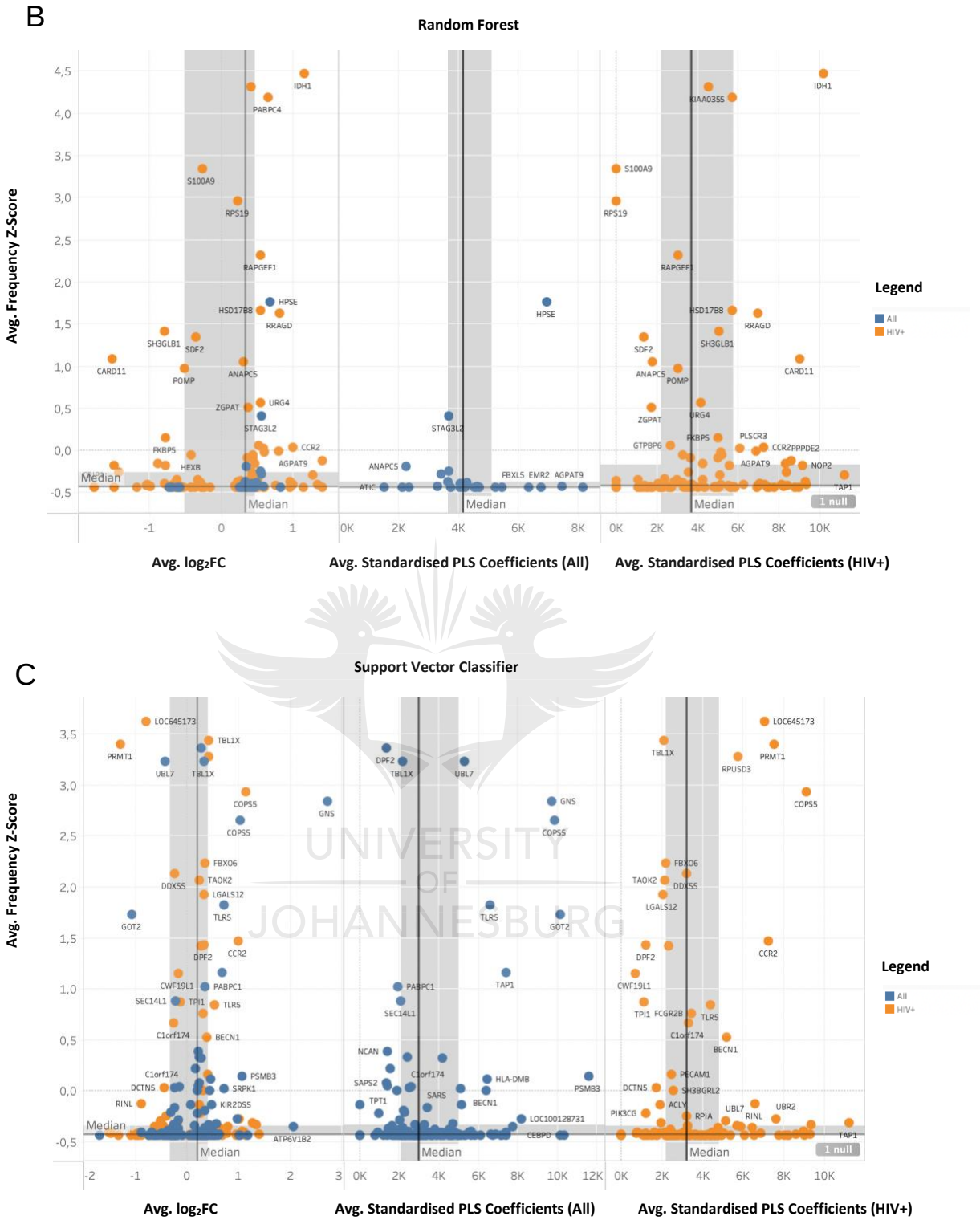


Figure 2.10 Cumulative frequency Z-scores of differentially expressed genes versus their average log$_2$ fold expression change (log$_2$FC) and standardised OPLS-DA regression coefficients (in HIV positive and 'all patients' groups.
Genes identified using the (A) Logistic Regression (LR) classifier (B) Random Forest (RF) classifier (C) Support Vector Classifier (SVC) are shown. Z-scores of the cumulative frequencies were used in order to cross-compare datasets. Genes with y-values lying in the upper quartile and x-values in the upper quartile (PLS regression coefficients) or outer quartiles (log$_2$FC) were considered significant.

Table 2.3 Top upregulated and downregulated DEGs in HIV positive patient group.
Transcript signatures for HIV positive patients, including the common gene name, log_2FC value, FDR value, direction of regulation and a description of the gene's function.

Group	***HIV Positive***				
Classifier	*Genes*	*log_2FC*	*FDR*	*Direction of Regulation*	*Description*
Logistic Regression/ Support Vector Classifier	CCR2	0.99194	1.33E-09	Up	The functional receptor for CCL2 and regulates the expression of T-cell inflammatory cytokines and T-cell differentiation.
	FCER1G	1.33557	6.38E-09	Up	An adapter protein containing an immunoreceptor tyrosine-based activation motif (ITAM) which transduces activation signals from various immunoreceptors.
	PRMT1	-1.28474	1.00E-08	Down	An arginine methyltransferase enzyme that methylates the guanidino nitrogens of arginyl residues present in protein substrates. PRMT1 mainly catalyses asymmetric dimethylation of histone H4 on arginine 3 (H4R3me2a), usually a marker of transcriptional activation, which has been implicated in transcriptional control
	RINL	-0.88937	7.25E-06	Down	Guanine nucleotide exchange factor (GEF) for RAB5A and RAB22A that activates RAB5A and RAB22A by exchanging bound GDP for free GTP. It plays a role in endocytosis via its role in activating Rab family members.
	UBR2	1.01844	2.04E-06	Up	E3 ubiquitin-protein ligase that recognises and binds to proteins bearing specific N-terminal residues that are destabilising according to the N-end rule, leading to their ubiquitination and subsequent degradation Plays a critical role in chromatin inactivation and chromosome-wide transcriptional silencing during meiosis via ubiquitination of histone H2A
Random Forest	AGPAT9/ GPAT3	0.79881	1.04E-07	Up	Converts glycerol-3-phosphate to 1-acyl-sn-glycerol-3-phosphate (lysophosphatidic acid or LPA) by incorporating an acyl moiety at the sn-1 position of the glycerol backbone.
	CARD11	-1.51021	5.28E-12	Down	Adapter protein that plays a key role in adaptive immune response by transducing the activation of NF-kappa-B (NF-κB) downstream of T cell receptor (TCR) and B-cell receptor (BCR) engagement.
	IDH1	1.15038	1.41E-12	Up	Provides instructions for making an enzyme called isocitrate dehydrogenase 1. This enzyme is primarily found in the fluid-filled space inside the cytoplasm. IDH1 converts a compound called isocitrate to another compound called 2-ketoglutarate. This reaction also produces a molecule called NADPH, which is necessary for many cellular processes.
	NOP2	-1.47994	3.65E-10	Down	S-adenosyl-L-methionine-dependent methyltransferase that specifically methylates the C_5 position of cytosine 4447 in 28S rRNA. May play a role in regulating the cell cycle and the increased nucleolar activity associated with cell proliferation.
	DESI1	0.81535	5.19E-08	Up	A chromatin enzyme, which enables SUMO-specific isopeptidase activity and identical protein binding activity. The enzyme is involved in protein desumoylation.
	RRAGD	-0.77986	2.30E-08	Up	Guanine nucleotide-binding protein that plays a crucial role in the cellular response to amino acid availability through regulation of the mTORC1 signalling cascade.

Table 2.4 Top upregulated and downregulated DEGs in 'all patients' group.
Transcript signatures for all patients, including the common gene name, log_2FC value, FDR value, direction of regulation and a description of the gene's function.

Group	***All***				
Classifier	*Genes*	*log_2FC*	*FDR*	*Direction of Regulation*	*Description*
Logistic Regression/ Support Vector Classifier	GNS	2.72251	3.00E-13	Up	Provides instructions for producing an enzyme called N-acetylglucosamine-6-sulfatase. GNS is involved in the step-wise breakdown of large molecules called glycosaminoglycans (GAGs)
	GOT2	-1.06568	3.52E-13	Down	Catalyses the irreversible transamination of the L-tryptophan metabolite L-kynurenine to form kynurenic acid (KA). As a member of the malate-aspartate shuttle, it has a key role in the intracellular NAD(H) redox balance. It is important for metabolite exchange between mitochondria and cytosol and amino acid metabolism. Facilitates cellular uptake of long-chain free fatty acids.
	PSMB3	1.06774	1.74E-10	Up	Non-catalytic component of the 20S core proteasome complex involved in the proteolytic degradation of most intracellular proteins. This complex plays numerous essential roles within the cell by associating with different regulatory particles. Two 19S regulatory particles form the 26S proteasome and thus participate in the ATP-dependent degradation of ubiquitinated proteins. The 26S proteasome plays a key role in maintaining protein homeostasis by removing misfolded or damaged proteins that could impair cellular functions and removing proteins whose functions are no longer required.
	TLR5	0.70875	9.65E-10	Up	Pattern recognition receptor (PRR) located on the cell surface activates innate immunity and inflammatory response. These receptors recognise distinct pathogen-associated molecular patterns that are expressed on infectious agents.
Random Forest	HPSE	0.684269	4.25E-11	Up	Endoglycosidase cleaves heparan sulfate proteoglycans (HSPGs) into heparan sulfate side chains and core proteoglycans. Participates in extracellular matrix (ECM) degradation and remodelling. Facilitates cell migration associated with metastasis, wound healing and inflammation.

2.4 Discussion

This study presented an OPLS-DA and ML pipeline to identify genes that are involved in the progression of latent to active TB in HIV positive individuals using existing microarray data. Integration of genes identified with log_2 FC values yielded an 11-gene minimal signature for latent TB (vs. active TB) in HIV positive individuals.

Below, we discuss the performance of the ML classifiers applied in this study and how they compare with one another. We then examine the observed functional redundancy in the gene expression patterns that distinguish latent TB from active TB in HIV positive individuals. The parallels and discrepancies between the results from this study and the original microarray study are also addressed.

2.4.1 General Performance of the ML classifiers and Comparison Between Models

In this study, the AUC was used to measure the overall discriminative performance of each ML classifier. Overall, the three classification models performed extremely well in predicting active TB and latent TB from expression data. Indeed, LR and SVC have been used in existing research for TB prognosis [143]. The LR model performed the best, followed by SVC and RF; however, these differences were not statistically significant. A similar observation had been reported by Abbas and El-Manzalawy [147] whereby the LR model outperformed RF, making it the preferred algorithm for developing prediction models based on gene expression profiles according to their work. Since the AUC values for the models generated were very similar, other performance metrics may be able to better discriminate between the performance of the models. Additional performance metrics that can be used in future for classification problems include specificity and sensitivity measures, logarithmic loss, or an F1 score.

Interestingly, the LR- and SVC-based methods yielded an identical list of genes. This similarity is most likely due to the algorithmic parallels between LR and SVM. SVC and LR are generalised linear models as they both create a decision boundary that linearly divides and classifies the data [180]. On the other hand, the RF classifier consists of a set of decision trees derived from a randomly selected subset of the training set. The votes from different decision trees are aggregated to determine the output predictions [141].

Previous research has established that linear decision functions can capture the underlying distributions in microarray classification tasks better than RF, suggesting that LR and SVC may be less sensitive to the choice of input parameters than RFs and can model linear decision functions more naturally than RFs [181]. In contrast, Uddin *et al.* [142] compared different supervised ML algorithms used for disease prediction. The study found that although the SVC algorithm is applied most frequently, the RF algorithm shows superior accuracy. Indeed, RF is currently one of the most widely used ML algorithms in TB detection [142, 144].

The different feature selection methods used in this study may additionally account for the differences observed between LR/SVC vs RF. In our study, the feature selection for the RF model used was based on Mean Decrease Impurity (MDI), while feature importance for LR and SVC was determined by comparing the standardised regression

coefficients of each input variable. These methods are the gold standards for feature selection in the respective algorithms. In the case of gene expression data, where the number of features or genes is large and the number of samples is small, it is common for the results to differ for each feature selection method [182]. Since the expression patterns and functions of all genes are unknown when applying a dataset, a useful approach in making the appropriate selection might be comprehensively evaluating the data using various feature selection methods. However, not all ML methods can be used to identify important features because their underlying methods are too complex to analyse the contributions of single covariates to the overall results [183]. This problem can be corrected by applying a bias-correcting measure of feature importance called 'permutation importance'. This method normalises the biased measure, returns significant p-values for each feature, and preserves the relations between features by using permutations of the outcome [183]. Permutation importance was not an option in this study due to computational constraints.

While the three different algorithms were very similar in terms of their ability to classify latent TB and active TB from the transcriptome dataset, they were not equal in the context of finding genes that discriminate between the two patient groups of interest (HIV positive vs 'all patients'). RF had the lowest Pearson *r* value comparing the two patient groups; here, RF was superior. We used genes from both LR/SVC and RF to define a transcriptomic gene signature of progression from latent to active TB in HIV+ individuals.

2.4.2 Comparison with the Original Microarray Study

The work by Kaforou *et a*l. primarily focused on defining a gene signature that distinguishes latent TB from active TB and active TB from other diseases (OD). Using an elastic net variable selection algorithm combining the Lasso and Ridge regression methods, they identified 27 transcripts that distinguished active TB patients from those with latent TB and 44 transcripts differentiating active TB from other diseases. Our integrative statistical and ML approach resulted in an 11-gene signature for HIV positive patients and a 16-gene signature for patients whose HIV status is unknown ('all patients' group).

We identified an overlap between the broader sets of genes generated from the OPLS-DA and ML algorithms and the 27 transcript signature from the study (a total of 8 and 12 overlapping genes for the HIV positive and 'all patients' groups, respectively).

However, our narrowed down 16-gene signature that distinguished latent TB from active TB for the 'all patients' group contained no genes in common with the 27 transcript signature. This is likely due to the use of different computational approaches in the two studies. This again points to the importance of applying various models to a dataset to ensure the inclusion of important genes that might otherwise have been disregarded if only one model were used. Another possibility for the lack of overlap between the gene signatures could be due to the exclusion of the 'other diseases' group in our analyses and the fact that we focused on HIV positive individuals.

The results of Kaforou *et al.* have been validated by several studies and provided the groundwork for further research. However, the genes involved in the progression of latent to active TB in HIV+ individuals remained unexplored. This study was able to reveal findings that can assist in addressing these missing elements.

2.4.3 Functional Redundancy Associated with Latent TB Genes of Interest

Following the ML and feature selection procedure, we observed several genes that appeared more frequently as top 10 features than others (Figure 2.6), suggesting the importance of these genes for distinguishing active TB and latent TB. Setting aside the distinct mechanics of the algorithms, it is interesting that the same genes did not always constitute the top 10 features. In the HIV+ group, 14, 17, and 92 genes were unique to LR, SVC and RF classifiers, respectively. While there is a random component to the machine learning algorithms used, the iterative approach followed would be expected to largely overcome any noise in the output; indeed, clear gene patterns emerged after 5000 iterations applied. It is possible that differences in the output between the algorithms employed reflected a redundancy in the gene expression patterns or processes involved in distinguishing latent TB from active TB in HIV positive individuals. This idea is supported by Chen *et al*., who noted that although ML methods have been proven to be successful in identifying disease-related genes, many methods fall short in considering the multifunction properties of many genes, particularly those that are associated with the disease [184].

ALPK1, PPPDE2, FCER1G and CCCR2 (Table 2.1) were genes of interest found in both groups investigated. The occurrence of these genes in both groups is likely due to the HIV positive group being a subset of the 'all patients' group. However, the degree of dysregulation of these genes differs between the two groups. To increase the discriminatory power of our final gene signatures, we excluded overlapping genes

found in both the HIV positive and 'all patients' cohorts. Additionally, some of the genes in our HIV positive latent TB signature have been implicated in the pathobiology of TB, thus supporting their significance as functional biomarkers of LTB infection in HIV+ individuals.

A number of genes and pathways found to be significant in this study have been linked to disease progression from latent TB to the TB disease state; however, it is unclear how the associated mechanisms and pathways occur in HIV infected patients. The following sections will discuss how some of these genes and pathways may be specifically relevant in HIV positive patients.

2.4.4 Biological Pathways, Interactions and Functions Associated with Latent TB, Active TB or TB/HIV coinfection which have been Studied in Literature

Immunosuppression in TB is a complex and not clearly understood phenomenon involving multiple mechanisms. An ideal biomarker should be capable of discriminating between latent TB and active TB and be functionally relevant. For this reason, examining the enriched pathways played an integral role in determining functionally relevant genes of interest in latent TB/HIV coinfected persons. Statistical and machine learning analysis of the transcriptomic dataset in this study identified DEGs enriched in biological pathways of the host immune response to *Mtb*, including CCR2, FCER1G and CARD11. Literature-based evidence has shown that the T cell receptor and B cell receptor signalling pathways are related to either active TB or latent TB infection. Lee *et al.,* for instance, reported T cell receptor (TCR) and B cell receptor (BCR) signalling pathways enriched by DEGs among latent TB and active TB groups [176]. The above pathways involving the identified genes are discussed below.

2.4.4.1 T cell receptor signalling pathway

Mtb regulates host immune response, primarily T cell responses, to survive, leading to disease or latent infection. This is mainly due to the organism living inside cells. As a result, T cells, rather than antibodies, are needed to eliminate the bacteria [177]. For proper T cell activation to occur, engagement with TCRs and antigen-presenting cells (APCs) in the presence of co-stimulation is required [177]. Any variation of this engagement could lead to T cell anergy (a tolerance mechanism whereby the lymphocyte is functionally inactivated) [177]. The pathogenic *Mtb* resides within

macrophages and inhibits several host cell procedures, which allows for its survival in the host cells [177]. However, the inhibited host processes and the molecules utilised by the pathogenic mycobacteria to accomplish intracellular survival are poorly understood [177]. Recent research suggests that TCR signalling is crucial for T cell memory and differentiation. The fate of T cell differentiation and how it is regulated has been extensively investigated [185]. However, there is limited understanding of T cell responses during the progression of latent TB to active TB, which inhibits the diagnosis of infection. IGRAs, which measure T cell responses to secreted protein antigens, have become the standard immunodiagnostic test of TB infection; however, these assays are poor predictors for the progression of latent TB to active TB [186].

Among the LTB HIV positive signature genes identified in our study, CCR2 participates in the T cell receptor signalling pathway. CCR2 is the functional receptor for CLL2, which regulates the expression of T cell inflammatory cytokines and T cell differentiation. Moreover, the expression profile of CCR2 has been shown to a play critical role in mediating alveolar macrophage migration during granuloma formation and has been described as a marker of terminally differentiated T cells [187]. However, the chemotactic signals promoting the recruitment of proinflammatory cytokines, including the CCR2 pathway, are unknown [187].

To understand the role of CCR2 in mediating cellular recruitment during *Mtb* infection, Riknink *et al.* infected CCR2 deficient mice with *Mtb* [188]. CCR2 deficient mice were more susceptible to *Mtb* infection and displayed increased pulmonary cellular composition due to a higher accumulation of neutrophils [188]. CCR2 was upregulated in latent TB and HIV positive individuals in the dataset used in this study. Thus, its expression decreases from latent to active disease state. This agrees with a recent study conducted by Guzman *et al*., which established that low expression of CCR2 in peripheral blood monocytes is a predictor of active TB and correlates with a high *Mtb* burden [187], although the HIV status of the subjects in their study was unknown. CCR2 acts as an entry coreceptor for HIV-1, and a mutation in the coding gene for the coreceptor, CCR2-64I, has been shown to delay disease progression of HIV-1 [189, 190]. This delayed progression is reflected in a slow CD4 T cell decline and in maintaining a stable viral load [189]. Further investigation of the modulation of CCR2 in the progression of LTB to active TB in HIV positive individuals and its regulation of T cell expression is needed.

2.4.4.2 B cell receptor signalling pathway

A series of studies have shown that B cells and antibodies may significantly reduce mycobacterial burden. With increasing evidence of B cells' ability to modulate immune response to *Mtb*, emphasis on characterising B cells might be of significant value. Older B cell knockout studies did not support the major role of B cells, but more recent studies have provided evidence that B cells and antibodies contribute to host defence against *Mtb* [191]. However, as with T cells, understanding the differential modulation of B cell responses during active and latent TB is limited.

Our study placed FCER1G in the 11-gene latent TB signature for HIV positive individuals. This adapter protein contains an immunoreceptor tyrosine-based activation motif (ITAM) which transduces activation signals from different immunoreceptors [192] It is a component of the Immunoglobulin E (IgE) receptor that plays a role in allergic reactions. Fc gamma receptors modulate immunity by engaging immunoglobulins (IgG) produced by B cells [192] and can potentially engage opsonising antibodies that protect against *Mtb* and thus impact mycobacterial survival [188]. A recent study reported FCER1G as one of the marker genes in *Mtb* infected alveolar macrophages in mice and human populations [193]. Even though limited studies exist linking FCER1G to latent *Mtb* infection, FCGR1A (a high-affinity receptor for the Fc region of immunoglobulin gamma receptor 1A) has frequently appeared as a transcriptional biomarker able to distinguish active TB from latent TB. The IntAct database indicates a molecular interaction between the two FCER1G and FCGR1A receptors. Kassa *et al*. also reported the potential of FCGR1A to discriminate latent TB from active TB in HIV patients [194]. In addition, research has reported increased FCGR1A expression in active TB patients compared to those that are latent TB infected [195]. TB treatment has been shown to significantly reduce its expression [195], suggesting the importance of FCGR1A in TB pathogenesis.

While our final gene signature did not contain FCGR1A, it did appear as an upregulated DEG in the 27 transcript signature of the original study conducted by Kaforou *et al*., albeit in the 'all patients' group. Given the abovementioned data and literature, FCGR1A is seemingly a key biomarker in LTB progression. A further understanding of its role in the B cell receptor signalling pathway will shed light on LTB/HIV coinfection. Further investigation of the FCGR1A gene might also unearth more knowledge on FCER1G since the two genes appear to be closely linked. Furthermore, FC gamma receptor gene polymorphisms, which influence receptor binding to IgG antibodies,

have been reported to likely play a critical role in the course of disease progression during HIV-1 infection [196]. These findings suggest that both the FCER1G and FCGR1A receptors might be involved in the progression of latent TB, particularly in individuals infected with HIV.

Furthermore, CARD11 was identified as a gene of interest in our study. This protein-coding gene plays a key role in the adaptive immune response by transducing the activation of NF-κB, JNK and mTOR downstream of the T cell receptor (TCR) and B cell receptor (BCR) engagement [197]. When T or B cells recognise a foreign substance, CARD11 is activated and binds to the BCL10 and MALT1 proteins, forming the CBM signalosome complex. The CBM complex proceeds to activate other protein complexes such as NF-κB, JNK and mTOR that are important for cellular signalling. NF-κB is a crucial transcription factor downstream of TLR that participates in a wide range of inflammatory diseases. Studies have highlighted evidence of how the inhibition of NF-κB activation affects the viability of intracellular *Mtb* in human macrophages [198]. It has been found that the inhibition of NF-κB activation decreases the viability of *Mtb* through the cellular processes apoptosis and autophagy, which are processes known to promote mycobacterial killing [198]. NF-κB, JNK and mTOR signalling direct the development of T cells and B cells to support immune response against foreign invaders [199]. The CARD11 gene might be a key modulator in preventing TB infection in both T cell and B cell signalling pathways. CARD11 has been identified as a frequently mutated gene associated with HIV-related diffuse large B cell lymphoma [200]. Still, the mechanism by which the gene can modulate B lymphocytes in HIV requires more comprehension. In addition, the CARD11 is required for NF-κB activation in T cells. Since canonical and noncanonical NF-κB pathways have been established in driving HIV expression [201], CARD11 could be involved in promoting HIV expression from latency.

It should be highlighted that the ‘signature’ genes reflect a minimal subset of genes that might be necessary for the classification of disease states, so there may be other genes closely correlated with CCR2, FCER1G and CARD11 that might provide insight into the biological processes driving the dysregulation exhibited.

The adaptive immune responses mediated by the abovementioned pathways are critical for the control of *Mtb*. Although, the manner in which T cell and B cell responses differ in persons with latent TB is not fully understood and warrants further exploration. Further understanding of the difference in TCR and BCR expression between active

TB and latent TB infection may be beneficial in the diagnosis and development of personalised treatment in subjects with *Mtb* [202]. This is also the case pertaining to HIV coinfection.

The fact that our study revealed genes that play a significant role in two pathways that have been extensively discussed in literature in relation to TB further supports the validity of the approach used to generate the gene signature. Using the associated genes for further downstream analyses, our research might fill the gap in understanding the mechanisms of T cell and B cell receptors in the progression of latent TB to active TB, particularly in HIV infected individuals. However, we recommend further studies, especially in sub-Saharan African patients, to confirm the role of these genes in latent TB/HIV coinfection and the associated biological pathways, which would establish their influence in latent TB/HIV coinfection therapies.

During our research, we have also revealed possible pathways who although no genes from our minimal gene signature were associated with these pathways, their possible involvement in the potential of the disease might require future investigation to be defined appropriately. Below we briefly describe the link between some of these pathways of interest with active or latent TB and HIV.

2.4.4.3 Positive regulation of cellular metabolic process

The ability of *Mtb* to recalibrate host metabolic processes in infected macrophages has been linked to its pathogenic success [203]. Several studies have illustrated the ability of *Mtb* in reprogramming macrophage metabolism, and it is believed that these adaptations might be crucial for its pathogenic success [203]. Metabolic changes induced by *Mtb* provide the necessary nutrients and could also rewire the activation state and the anti-microbial effector functions of infected macrophages [203]. Over the recent years, although still unclear, many studies in the emerging field of immunometabolism have attempted to describe the associations between macrophage metabolic states along with their immunological responses [203]. In the context of LTB infection, latent *Mtb* are less metabolically active and have diminished replication rates compared to bacilli in active TB disease [20]. Emerging evidence indicates the vital role of metabolic pathways usage in immune cells in HIV-1 pathogenesis [204]. Immunometabolism shapes immune responses against infection as cell metabolic products are key drivers of inflammation [204]. Moreover, the metabolic pathways of CD4+ T cells determine their susceptibility to HIV-1 infection

and the persistence of these infected cells [204]. However, further knowledge is required to understand the links between cellular metabolic processes and latent TB/HIV coinfection.

2.4.4.4 Insulin signalling pathway

Although immunity to TB in the lung and lymphoid system has been studied intensively, very little is known about the involvement of adipose tissue and non-immune cells in the interaction between the host and pathogen during the disease [160]. Using a mouse model infected with *Mtb*, Martinez *et al.* found TB to be associated with increased insulin signalling and systemic glucose tolerance in adipocytes [160]. TB infection promoted adipose tissue inflammations and adipocyte hypertrophy, both conditions typically associated with insulin resistance [160]. The synergic association between diabetes and TB has recently emerged as a global health concern due to the increasing prevalence of diabetes in TB endemic regions [153]. Yoo *et al*. investigated the association of diabetes status with risk of TB incidence and found that individuals with diabetes had a 48% higher risk of TB incidence than individuals without diabetes [160]. These findings support the positive association between TB risk and diabetes duration [160]. The study also confirmed that reduced insulin receptor expression and downstream signalling in T lymphocytes are found in patients with diabetes [160]. Since diminished insulin signalling includes proinflammatory cytokine production in T cells and antigen-specific proliferation, reduced insulin production and altered receptor-mediated signalling in T cells could result in increased TB risks in patients with diabetes [153]. Research has shown insulin resistance contributes to the metabolic alteration observed in HIV positive patients [205]. In HIV untreated patients, there is severe insulin resistance with increased LPS and cytokines that involves adipose tissue, love, hypothalamus, vessels and muscle [205, 206]. While HIV patients that undergo antiretroviral drug therapy show mild/moderate insulin resistance with reduced LPS and cytokines, although there is a decrease in proinflammatory cytokines, they do not completely return to normal, indicating some level of inflammation that persists [205]. As such, the possible role of the insulin signalling pathway in the progression of TB infection within HIV positive groups is an avenue that might need to be explored.

2.4.4.5 Endocrine resistance

Although TB is typically seen as a pulmonary disease, extrapulmonary TB, which affects various organs and other systems, is not unusual. Endocrine gland

involvement in TB has markedly diminished due to the availability of effective anti-TB therapy [207]. Although rare, TB of endocrine glands (such as adrenals, thyroid and pituitary) has pathophysiological effects that have been established to be clinically significant [207]. Other endocrine glands that might be affected by TB include the thymus, pancreas, pineal gland, parathyroid and gonads. Furthermore, research has confirmed that even in the absence of gland involvement, the TB disease process and anti-TB treatment could result in endocrine and metabolic derangements [207]. HIV infected patients have a high risk of developing endocrine disorders. The endocrine glands are affected in various ways, such as functional derangement, resultant immune suppression, invasion of neoplasms, and opportunistic infections [208]. Some endocrine abnormalities associated with HIV include growth hormone deficiencies, which may also contribute to insulin resistance, and growth hormone resistance.

Along with other glandular dysfunctions such as hypopituitarism, thyroid disorders and hypogonadism [209]. The most affected gland in HIV is the adrenal gland [209]. A deeper understanding of endocrine resistance in latent TB infection is required to establish how the pathway might affect disease progression in immunocompromised HIV patients.

2.4.4.6 ATP binding

Mtb is able to survive low-energy conditions. This allows for infections to remain dormant and thus decreasing their susceptibility to many antibiotics [210]. Bedaquiline, a novel therapeutic drug used to treat multi-drug resistant tuberculosis, can sterilise even LTB infection [210]. The drug works by targeting the mycobacterial ATP synthase; an enzyme reported to be essential in *Mtb* for optimal growth. However, the manner in which the drug binds the intact enzyme is unknown [210]. *Mtb*'s ability to persist in the latent state has been associated with the pathogen's ability to adapt to host induced metabolic constraints such as oxygen stress [20]. The production of ATP is much greater in the presence of oxygen; thus, to survive in an oxygen-deprived environment, the bacteria need to alter their metabolic pathways to depend on anaerobic respiration or find alternative mechanisms to generate energy [20]. These survival mechanisms of the pathogen are potential contributors to latent TB progression. The HIV-1 accessory protein Nef is essential for viral replication, and disease progression and studies have shown that Nef mediates functional impairment of ATP binding cassette transporter A1 (ABCA1) and suppresses cholesterol efflux

[211]. The Nef-mediated inactivation of ABCA1 results in an accumulation of cholesterol in macrophages, the increase in abundance of lipid rafts and elevation in cholesterol content of viral membranes [211]. These effects consequently increase HIV production and infectivity. As such, further exploration of the ATP binding pathway's involvement in latent TB progression within the context of HIV is needed.

2.5 Conclusion

By means of a computational, integrative approach that leveraged statistical DE analysis and ML feature selection we were able to determine a broader list of genes of interest and a minimal gene signature for latent TB in HIV positive persons. A network enrichment analysis and literature search was applied to these genes of interest to elucidate the biological pathways associated with latent TB/HIV coinfection. A number of these observed genes were linked to biological processes that require further investigation to establish these pathways possible involvement in latent TB progression in immunocompromised individuals.

3 PREDICTION OF EPIGENETIC MECHANISMS INVOLVED IN PROGRESSION FROM LATENT TO ACTIVE TB IN HIV POSITIVE INDIVIDUALS

3.1 Introduction

Several studies indicate that susceptibility or resistance to active TB disease goes beyond genetic influences encoded in the DNA and that disease risk may be influenced by epigenetic variation [212]. Epigenetic mechanisms play a pivotal role in regulating gene expression during cellular response to extracellular stimuli [125]. Epigenetic regulation involves a combination of various molecular and biochemical mechanisms, including transcription factor (TF) binding, histone-modifications (HM), DNA methylation and non-coding RNA. The epigenetic regulation of transcriptional profiles in TB disease is poorly understood and much less so in latent TB infection. The identification of epigenetic alterations associated with latent TB could be used as targets in therapies that are aimed at reducing the systematic activation state in HIV infected patients.

High-throughput sequencing analyses such as chromatin immunoprecipitation sequencing (ChIP-seq), RNA sequencing (RNA-seq), and DNA affinity purification sequencing (DAP-seq) typically produce sets of genes of interest requiring further analyses to ascertain their underlying regulatory mechanisms and biological implication [213]. As such, it is important to focus on sets of genes that share biologically important attributes. Enrichment analysis can support the discovery of biological functions which may have been missed in a resultant gene set [214]. Enrichment analysis is undertaken on a gene set of interest identified using high throughput genomic methods to provide insight into the biological function underlying a list of genes [214]. The analysis maps genes and proteins to their associated biological annotations and compares them with all genes represented on a microarray chip [214]. Enriched terms or marks are defined as those that are statistically over- or underrepresented within the gene list [215]. An *in-silico* approach can be performed to characterise the epigenetic marks of genes of interest that are likely modulated in TB infection. The annotations and retrieval of enriched epigenetic features of a gene list can be done using web-based tools that utilise sequencing data from reference experimental databases.

A bioinformatics tool called AnnoMiner has a transcription factor (TF) and histone modification (HM) enrichment analysis function, which identifies enriched peaks in the promoter regions of a user-provided gene list [216]. This gene list could be one generated from a transcriptomic analysis. AnnoMiner performs TF and HM enrichment analysis using ChIP-seq datasets taken from the ENCODE, modENCODE and modERN databases [216]. The tool considers a gene as a potential target if its promoter overlaps with a TF or HM peak. AnnoMiner's enrichment function provides a 'dynamic ranges' option that automatically detects the optimal threshold to define the up or downstream boundary for each TF and HM. Alternatively, the user can manually define the promoter region up or downstream of the annotated transcription start site (TSS) of the genes of interest. Additionally the user can also set the minimum required overlap, in base pairs (bp) or percentage, to consider binding biologically relevant [216].

Considering that a single miRNA is able to target multiple genes and a single gene can simultaneously be targeted by more than one miRNA, it is crucial to narrow down a large list of miRNA-target interactions to gain insights into the mechanisms regulated by miRNAs in a variety of cellular processes. Over the years, numerous bioinformatics tools related to miRNAs have been established to predict candidate mRNAs based on information related to the sequence and evolutionary conservation [217, 218]. The major drawback of these bioinformatics methods is that they typically result in the prediction of tens or hundreds of targets for each miRNA, usually with high false-positive rates [219]. Consequently, further experiments are required to determine which of the predicted targets are genuinely targeted by miRNAs. This is usually hindered by the unfeasibility to experimentally validate all candidate genes individually [220]. A solution to this issue is filtering out those that are statistically insignificant, prioritising miRNA target interactions, and investigating these candidate interactions more thoroughly [220]. Although computational methods that tackle this prioritisation and validation problem exist, they still require navigating multiple websites and merging the results for further analysis [220].

The interactive web tool MIENTURNET (MicroRNA Enrichment TURned NETwork) performs a miRNA target enrichment analysis using an input list of genes. The tool retrieves data of experimentally validated and computationally predicted miRNA-target interactions from the miRTarBase and TargetScan databases, respectively [220]. It then filters based on statistical significance resulting from a miRNA target enrichment analysis [220]. The miRTarBase databases report miRNA target interactions that have

been experimentally validated by microarrays, reporter assay, western blot and next-generation sequencing experiments [220]. Furthermore, MIENTURNET captures topological properties of the miRNA regulatory network, which would not be apparent through the pairwise analyses of individual components [220].

Using the identified minimal transcriptomic gene signature of latent TB in HIV infected individuals (Chapter 2), we performed an enrichment analysis to study the epigenomic landscape of different cell lines to determine the epigenetic mechanisms that may be relevant to latent TB during HIV infection using enrichment analysis. A selection of suitable cell lines was made for our specific research question for the epigenetic enrichment analyses. We selected the major innate immune cell types involved in TB infection: PBMCs, T cells, and B cells. Enrichment analyses are typically performed by testing for TF and HM overrepresentation in the promoter regions of user-provided gene lists. Our epigenomic analysis focused on identifying enriched TFs, HMs and miRNAs.

3.1.1 Aim

To predict possible epigenetic mechanisms that underly the regulation of genes involved in latent to active TB disease progression in HIV positive individuals using enrichment analysis.

3.1.2 Objectives

The specific objectives include:

- Identifying transcription factors and histone modifications that may punctuate latent to active TB disease progression through AnnoMiner enrichment analysis.
- Identify miRNAs that may regulate genes involved in latent to active TB disease using the MIENTURNET enrichment analysis tool.

3.2 Methods

3.2.1 Transcription Factor and Histone Modification Enrichment Analysis

A transcription factor analysis was conducted with the AnnoMiner web server. Two separate plain .txt files containing a list of gene names were used as an input, including the 11-gene latent TB signature for the HIV positive cohort and the 5-gene signature identified for the 'all patients' group. Since AnnoMiner requires a minimum query size of ten genes, an additional five genes (POLB, KIF22, MEF2D, EVL, PVRIG, ZNF438)

from our broader OPLS-DA and ML analysis were added to the 'all patients' group. However, these five additional genes were not considered in the final analysis. The *Homo sapiens* (hg38) reference genome was used and GENCODE was the reference human genome annotation of choice for the analysis. The default minimum overlap of 1 bp was selected to define biological relevance and promoter regions were defined as 500 bp downstream and 1000 bp upstream to the TSS. These promoter boundaries were defined based on the recommended regions suggested by Georgakilas *et al.* [221]. A results table containing experiment information, lists of target genes, enrichment scores (a measure of magnitude of enrichment) and the FDR *p*-values were obtained along with a bar plot showing the first 10 top-ranking results ranked by combined score (defined as the score of the hypergeometric test *-log10(*p*-value)).TFs or HMs were sorted by ascending order of FDR adjusted *p*-values (FDR *p*-values > 0.05 were excluded). Only TFs and HMs enriched in cell lines of interest (PBMCs, B cells and T cells) were considered. A literature search was conducted to investigate the role of these TFs or HMs in latent and active TB regulation.

3.2.2 MicroRNA Enrichment Analysis

MIENTURNET was utilised for miRNA target enrichment analysis. The .txt files containing the lists of gene names, including the 11-gene latent TB signature for the HIV positive cohort and the 5-gene signature identified for the 'all patients' group were uploaded to the tool. As with the TF and HM analysis, an additional five genes (POLB, KIF22, MEF2D, EVL, PVRIG, ZNF438) from our broader OPLS-DA and ML analysis were added to the 'all patients' group gene set. However, these five additional genes were not considered in the final analysis. The default threshold value of 2 for the minimum number of miRNA target interactions was selected. TargetScan and miRTarBase reference databases were used for the enrichment analysis. The results table, including target genes, *p*-values, FDR-adjusted *p*-values, odd ratios, and the number of interactions, was downloaded as a CSV file along with bar plots showing the top 10 target genes resulting from the enrichment analysis. Only the miRNAs reported as statistically significant (FDR p-value < 0.05) were assessed. A literature review of the miRNAs of interest was conducted to determine their epigenetic modulation in latent or active TB.

3.3 Results

3.3.1 Transcription Factor Enrichment Analysis

To predict the involvement of transcription factor binding in the regulation of latent TB associated genes of interest in both the HIV positive and the 'all patients' group, an enrichment analysis was performed to identify TFs in the defined promoter region of the selected genes. Figure 3.1 illustrates the most significantly enriched TFs in both patient groups (the full results table can be found in Supplementary Table S5). The HIV positive cohort contained higher combined scores for the top 10 hits, indicating higher TF binding densities for each individual TF compared to the combined scores of the 'all patients' group. Only the enriched TFs found within the cell lines of interest (PBMCs, B cells and T cells) were selected for further investigation. Of these, YY1, SRSF3 and ATF3 (in the HIV positive group) have been reported in literature in relation to latent or active TB (Table 3.1). One enriched TF, namely KDM1A, in the HIV positive cohort whose involvement has not previously been reported in TB research may be an additional candidate for epigenetic regulation of TB disease progression, based on its modulation and target genes.

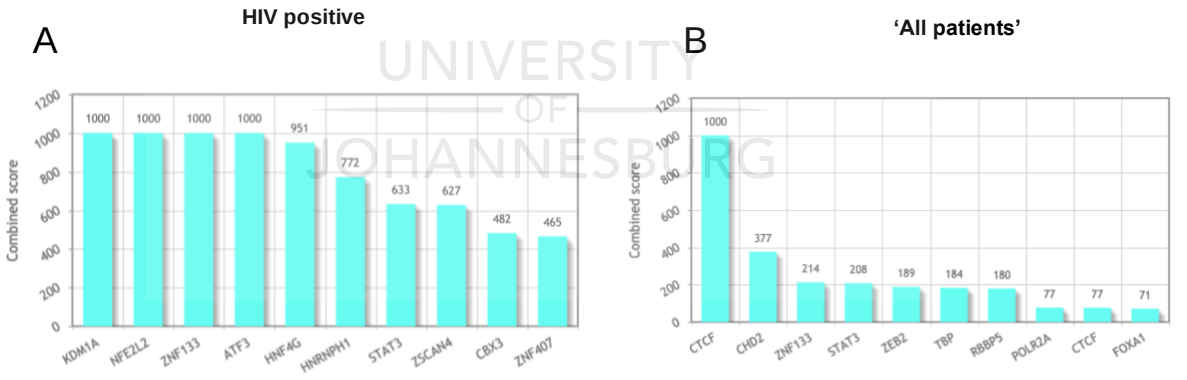


Figure 3.1 The 10 most enriched TFs found in the promoter regions of the minimal latent TB signature target genes within the HIV positive and (B) 'all patients' groups.
The identity and combined score of the enriched TF are shown. The combined score was calculated by multiplying the -$\log_{10}$(p-value) to the score of the hypergeometric test (computed as (List Hist/List Size)/(GenomeHits/Genome Size)).

Table 3.1 The top enriched TFs in PBMCs and immune cells identified using AnnoMiner in HIV positive patients, which have been reported in the context of latent TB or active TB in literature. Only enriched TFs with an FDR < 0.05 were considered statistically significant.

	Top Enriched TFs	*Literature report in the context of latent or active TB*	*Sample*	*Cell-Line*	*FDR*	*Target Gene*
HIV Positive	YY1	YY1 is known to play a crucial role in the maintenance and progression of some pulmonary diseases; however, its specific role in TB remains unknown. Recent studies in a mouse model have elucidated the role of YY1 in regulating the transcription of CCL4. YY1, CCL4 and TGF-β were found to be overexpressed in the lung tissue of TB infected mice during the late stage of the disease. YY1 is overexpressed in experimental and human TB. Thus treatment that decreases YY1 expression may be a new therapeutic strategy against TB [222].	GM12891	B-Lymphocyte	3.60E-07	IDH1;RRAGD
			K562	Human blood (chronic myelogenous leukemia)	2.17E-04	GPAT3
	SRSF3	The expression of genes encoding SR proteins in *Mtb* has been evaluated. SRSF2 and SRSF3 were found to be significantly downregulated post *Mtb* infection. These findings suggest that alternative splicing might be involved in host gene regulation during *Mtb* infection of macrophage cells [223].	K562	Human blood (chronic myelogenous leukemia)	9.02E-05	PRMT1
	ATF3	TF activating transcription factor 3 (ATF3) is shown to be upregulated during early infection of macrophages in mice. ATF3 depletion promotes mycobacterial survival in macrophages suggesting its protective role in the host [224].	K562	Human blood (chronic myelogenous leukemia)	7.25E-07	GPAT3
	KDM1A*	KDM1A demethylates H3K4me1/2, and together with the histone deacetylases HDAC1/2, it forms part of co-repressor complexes recruited by zinc finger factors to control transcription [225]	GM12878	B-Lymphocyte	2.98E-08	PRMT1

**Refers to TF that has not been directly reported in TB related literature but might be of interest in future studies*

3.3.2 Histone Modifications Enrichment Analysis

To broadly assess the histone modifications associated with regulating the genes of interest, an HM enrichment analysis was performed. Figure 3.2 depicts the ten most significantly enriched HMs in both patient groups (the full results of enrichment hits across various human cell types and tissues can be found in Supplementary Table S6). Like the TF enrichment analysis, only the enriched HMs found within the cell lines of interest (PBMCs, B cells and T cells) were investigated. Three histone marks (H4K20me1, H3K27me3 and H3K4me1) have been associated with latent TB or active TB in literature, while another three HMs (H3F3A, H3K36me3, H3K27ac) have not been reported in literature related to TB and require further investigation.

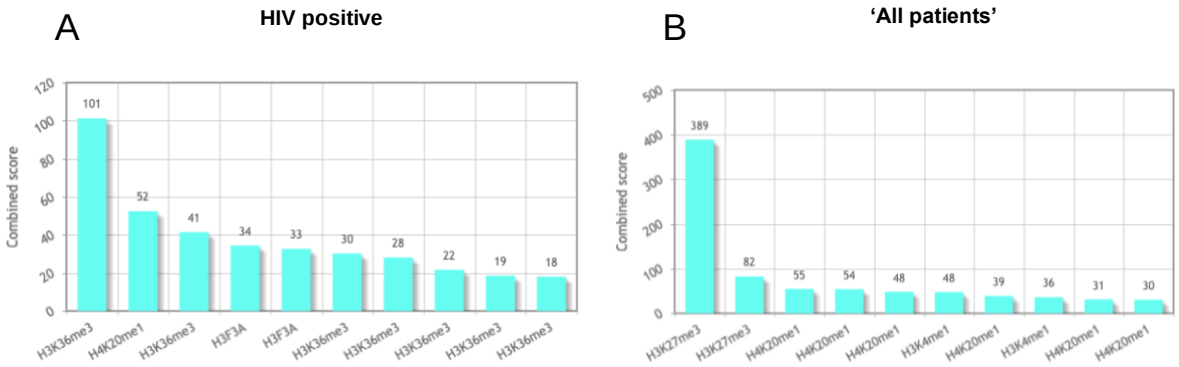


Figure 3.2 The 10 most enriched HMs found in the promoter regions of the signature target genes within the (A) HIV positive group and (B) 'all patients' group.
The identity and combined score of the enriched HM are shown. The TFs have been ranked by their combined score. The combined score is calculated by multiplying the $-\log_{10}$(p-value) to the score of the hypergeometric test (computed as (List Hist/List Size)/(GenomeHits/Genome Size)).

Table 3.2 The top enriched HMs in immune cells determined using AnnoMiner in HIV positive which have been reported in the context of latent TB or active TB in literature.
Only enriched HMs with an FDR < 0.05 were considered statistically significant.

	Top Enriched Histone Modifications	*Functional Association*	*Literature report in the context of latent or active TB*	*Sample*	*Cell Line*	*FDR*	*Target Gene*
HIV Positive	H4K20me1	Transcriptional activation	It has been reported that SET8, a histone H4 lysine 20 monomethylase (H4K20me1), is highly induced during *Mtb* infection. The epigenetic reprogramming of the host cell by the SET methyltransferase promotes the survival of *Mtb* in macrophages through the regulation of apoptosis and inflammation. SET8 can orchestrate immune evasion strategies by initiating NQO1 and TRXR1 and regulating the *Mtb* induced expression of these two reductases. The loss-of-function studies in a TB mouse model support the critical role of SET8-NQO1/TRXR1 in *Mtb* survival. Therefore, enhancing host immune responses against *Mtb* by harnessing SET8-NQO1/TRXR1 with its specific and potent inhibitors could lead to host-directed therapeutic adjuvants for TB treatment [226].	Loucy	T-cell leukemia	3.77E-02	DESI1; IDH1; PRMT1

Table 3.3 The top enriched HMs in immune cells determined using AnnoMiner in HIV positive and all patients, which have not been directly reported in TB related literature.
Only enriched HMs with an FDR < 0.05 were considered statistically significant.

	Top Enriched Histone Modifications	*Functional Association*	*Literature report*	*Sample*	*Cell Line*	*FDR*	*Target Gene*
HIV Positive	H3F3A	Transcriptional activation	H3F3A's target gene, PRMT1, enhances AKT signalling by methylating Er-alpha. This pathway plays a protective role in *Mtb* infection and is targeted by the pathogen to evade the host immune system by modulating the host defence mechanisms [227]	NCI-H929	B-lymphocyte	5.00E-03	FCER1G; PRMT1
				MM.1S	Peripheral blood of a multiple myeloma patient	3.15E-02	FCER1G; IDH1; PRMT1
	H3K36me3	Transcriptional activation	Evidence indicates that pathogens such as *Mtb* can alter DNA methylation and regulate the function and expression of DNA methylation modifiers such as DNMTs. It has been reported that a gene body enriched with H3K36 tri-methylation (H3K36me3) or H3K9me3 is favourable for DNMT3B recruitment, resulting in hypermethylation at these regions that functionally relate to gene transcription initiation, proper splicing and compact chromatin at active genes [228].	DND-41	T-cell leukemia	1.52E-02	DESI1; FCER1G; IDH1; PRMT1
				DOHH2	B-cell lymphoma	1.22E-02	DESI1; FCER1G; IDH1; PRMT1
				KOPT-K1	T-cell acute lymphoblastic leukemia	2.69E-04	FCER1G; IDH1

3.3.3 *miRNA Target Enrichment Analysis*

One of the main epigenetic mechanisms includes regulation by non-coding RNAs. The aberrant expression of miRNAs, in particular, can alter the DNA or chromatin state by restricting chromatin remodelling enzyme activity. Because miRNAs have been well understood as epigenetic modulators and can be modulated by epigenetic changes, a miRNA target enrichment analysis was executed using the MIENTURNET web tool. We identified hsa-miR-3135a as the only significantly enriched miRNA in the HIV positive group, with an FDR value of 0.05738 (Figure 3.3). This conserved miRNA was retrieved from the miRTarBase database containing experimentally validated miRNA target interactions in various human cell lines. The identified miRNA's target genes include desumoylating isopeptidase 1 (DESI1) and protein arginine methyltranserase 1 (PRMT1). No significantly enriched miRNAs were found in the 'all patients' group (Figure 3.4.).

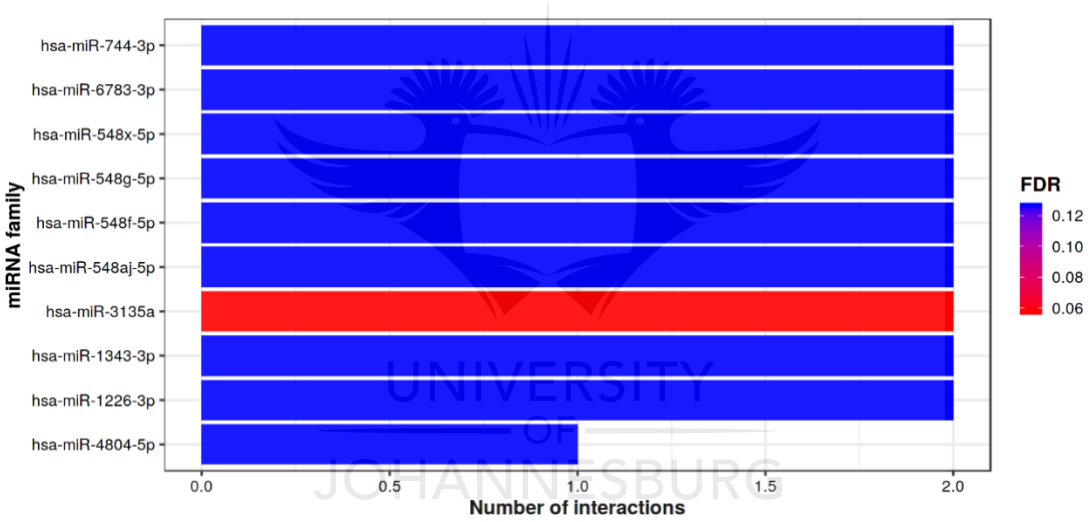


Figure 3.3 miRNA enrichment analysis using MIENTURNET (taken from miRTarBase) showing the 10 top miRNA families enriched in the HIV positive group and their number of validated target interactions.
The colour code reflects the increasing FDR value. miRNAs with FDR < 0.05 were considered statistically significant.

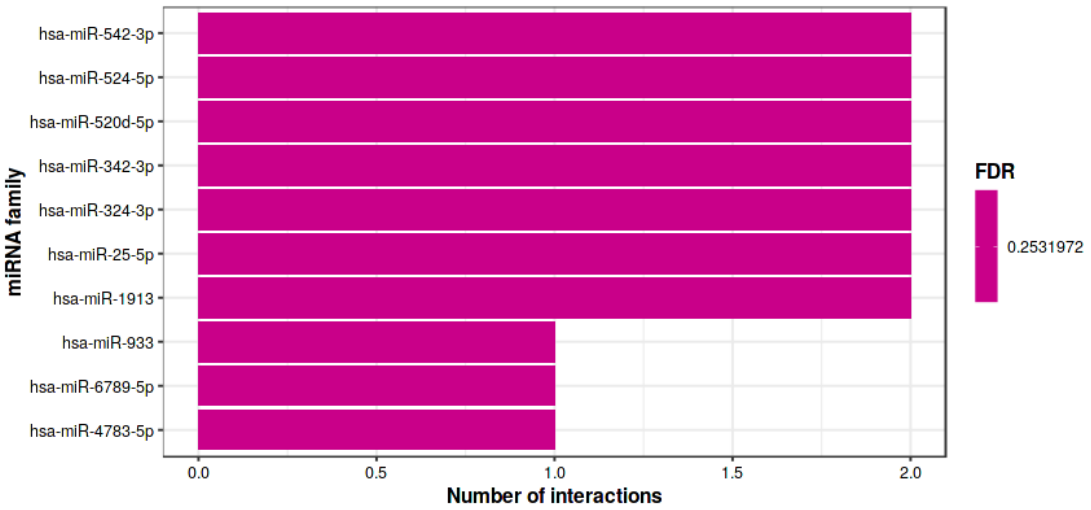


Figure 3.4 miRNA enrichment analysis using MIENTURNET showing the 10 top miRNA families enriched in the 'all patients' group and their number of validated target interactions.
miRNAs with FDR < 0.05 were considered statistically significant.

3.4 Discussion

Recent reports have highlighted the influence of pathogenic *Mtb* in modulating the transcriptional profile of host defense associated genes by influencing epigenetic factors. *Mtb* infection can alter the host epigenome to modulate the transcriptional machinery and thus trigger susceptibility to disease. This mechanism of epigenetic alterations during latent TB infection is not fully understood.

The aim of the work presented in this Chapter was to identify epigenetic mechanisms that may play a role in the regulation of genes linked to the progression from latent to active TB in HIV positive individuals through enrichment analyses. Accordingly, several TFs, HMs and miRNA enriched in blood and immune cells of available ChIP-seq data are of interest. A comprehensive list of TFs and HMs found in other cell lines can be found in the supplementary section (Supplementary Table S5-S7). While experimental data in HIV infected models are lacking, the enriched TFs and HMs identified in PBMCs, B- and T-cells, may provide a focus for future efforts to uncover the latent TB epigenetic landscape in HIV positive individuals and its role in the host immune response.

3.4.1 Transcription Factors that may regulate progression from latent to active TB in HIV positive individuals

We identified three TFs (YY1, SRSF3 and ATF3) enriched in the promoters of the genes of interest (HIV positive group) in PBMCs and immune cells, that have been linked to latent and active TB in literature.

Yin-Yang-1 (YY1) contributes to the maintenance and progression of some pulmonary diseases, including pulmonary fibrosis [222]. Although, the role of YY1 in TB remains unknown. Research conducted by Santiago *et al.* aimed to elucidate the role of YY1 in the regulation of the CCL4 chemokine and its implication in TB [222]. The study aimed to determine whether YY1 regulates CCL4 using reporter plasmids, ChIP and siRNA assays and measuring the expression of YY1 and CCL4 in a mouse model of TB [222]. Their results indicated that YY1 regulates the transcription of CCL4. Moreover, YY1 and CCL4 were overexpressed in the lung tissues of mice with TB during the late stages of the disease and the tissues of TB patients [222]. This could reflect the possibility of an increase in the expression of YY1 during the progression of latent to active TB. Therefore, treatments that decrease YY1 expression may be examined as a therapeutic strategy against latent TB progression.

YY1 is a transcriptional activator or repressor depending on the context and is known to interact with chromatin modifiers, suggesting that chromatin modifications may determine the direction of regulation [229]. YY1 functions as a Polycomb group (PcG) protein and initiates methylation of histone 3 lysine 27 (H3K27me3). While H3K27me3 had the highest combined score (Figure 3.2) in the 'all patients' group, this modification was not enriched in the promoters of genes linked to latent TB in HIV positive patients in our analysis. Furthermore, YY1 promotes transcription through the recruitment of the PRMT1 methyltransferase [229]. Indeed, PRMT1 has been identified in this study as a gene of interest in the HIV positive group. PRMT1 is an arginine methyltransferase known to methylate histone 3 arginine 3 (H3R3me2a), a transcriptional activation marker [229]. However, H3R3me2a enrichment in the promoters of the genes of interest was not observed in PBMCs and immune cells of this study.

IDH1 and AGPAT9, identified as genes of interest in this study, are target genes of YY1. IDH1 produces NADPH, a molecule necessary for many cellular processes [230]. NADPH oxidases play critical roles in antimicrobial host defense and inflammation beyond the production of reactive oxygen species (ROS) [231]. A well-studied mechanism whereby HIV infection is shown to directly impact oxidative stress is the interaction between the NADPH-oxidases and the HIV protein tat [232]. Tat (Trans-Activator of Transcription) is a regulatory protein encoded for by the Tat gene in HIV-1 which drastically enhances the efficiency of viral transcription [233]. The HIV-1 tat protein activates Rac1 which is crucial for the activation of various isoforms pf NADPH-oxidases [234]. AGPAT9 catalyses the conversion of glycerol-3-phosphate to

lysophosphatidic acid during the synthesis of triacylglycerol. Pathogens may use stored lipids inside host cells as a potential energy source, particularly under serum-starved conditions [230]. An important study by Raja *et al.* reported that serum starvation leads to the reactivation of HIV-1 latency in monocytes through the ERK/JNK pathway [235]. It is, therefore, possible that AGPAT9 supports mycobacterial survival and growth under conditions that may be unique in HIV positive individuals. While experimental evidence is lacking, these connections present interesting leads for future investigation.

The RNA-binding SRSF3 protein is a splicing factor and regulator of pre-mRNA alternative splicing [236]. Zhang *et al.* found the protein to be significantly downregulated post-*Mtb* infection [223]; however, its regulation during latent TB remains unclear. SRSF3, along with other members of the SR protein family (such as SRSF2, SRSF3 and SRSF), is a strong repressor and can significantly downregulate Tat activity [237]. The role of SRSF3 in *Mtb* progression in HIV positive individuals needs to be elucidated in future studies. A meta-analysis study conducted by Chen *et al.* identified ATF3 as one of four most statistically significant genes in combined TB infection with HIV positive patients [149]. ATF3 was found to be upregulated in active TB and HIV positive infected individuals compared to the control group (TB negative and HIV positive). Further studies possibly exploring the role of the ATF3 mark in latent TB might unearth more knowledge regarding the mechanism involved in progression to the active disease state during HIV infection.

The analyses also identified an enriched TF, namely KDM1A, that has not previously been reported in TB disease progression *per se* but has some links with *Mtb* and HIV infection. KDM1A was the first demethylase to dispute the concept of the irreversible nature of methylation marks and has emerged as an epigenetic developmental regulator [238]. The lysine demethylase has been reported to cooperate with CTIP2 to repress HIV-1 transcription and viral expression [239]. To the best of our knowledge, no literature discusses KDM1A's role in *Mtb* infection.

3.4.2 Histone Modification Marks that may regulate TB disease progression in HIV infected individuals

Like the TF enrichment, the selected HMs reported in this study are those of immune cell lines (both healthy and diseased cells), which do not represent HIV infected cell lines as ChIP data on HIV infected cells was not available. Based on literature, we identified one histone mark, H4K20me1, from the HIV positive group that has been

associated with latent TB or active TB and an additional two HMs, namely H3F3A and H3K36me3 in HIV infected patients that have yet to be reported in TB related literature and require further investigation.

Singh *et al*. analysed the SET8 methyltransferase associated with H4K20me1 to define the epigenetic regulation of inflammation during *Mtb* infection through the induction of NQO1 and TRXR1 [226]. They found that SET8 mediates H4K20me1 modification on *Mtb*-triggered promoters of NQO1 and TRXR, regulating inflammation and apoptosis, and thereby assisting *Mtb* survival [226]. The promotors of DESI1, IDH1 and PRMT1, identified in this study as genes of interest, are enriched with H4K20me1 marks in PBMCs and immune cells. These genes may therefore be regulated by this modification. H4K20me1 is associated with transcriptional activation [240]; indeed, our analysis showed that DESI1 and IDH1 are upregulated in latent TB/HIV coinfected individuals. In contrast, PRMT1 in our study is downregulated in these individuals, suggesting that PRMT1 promoters are less extensively marked by H4K20me1 than DESI1 and IDH1.

Previous studies have indicated that Rv1988, a functional methyltransferase, share similarities with PRMTs [241]. Decreased levels of Rv1988 in *Mtb* reduce bacterial survival in the host through epigenetic control of host cell transcription [241]. As such, it is possible that reduced expression of PRMT1 could inhibit *Mtb* survival and consequently prevent the progression from latent to active TB. These genes and their association with the H4K20me1 mark need to be further investigated to explore the epigenetic regulation that occurs during *Mtb* and HIV coinfection. Aditionally, PRMT1, which is a target gene of H3F3A, enhances AKT signalling by methylating Estrogen receptor alpha. This pathway plays a protective role in *Mtb* infection and is targeted by the pathogen to evade the host immune system by modulating the host defence mechanisms.

Mtb has been reported to change DNA methylation [228]. DNA methyltransferase genes such as DNMT3A, DNMT3B and DNMT3L have been found to have differentially methylated regions (DMRs) related to *Mtb* infection [242]. Qin *et al.* reported that enrichment of H3K36me3 and H3K9me3 marks in gene bodies is favourable for DNMT3B recruitment, resulting in DNA hypermethylation at the regions functionally relating to the initiation of gene transcription, splicing and compact chromatin at repressed genes [228]. Exploring the role of the H3K36me3 mark might reveal crucial knowledge in the progression of TB infection in HIV infected persons.

3.4.3 *Possible regulation of TB disease progression in HIV positive individuals by hsa-miR-3135a*

miRNAs can induce chromatin remodelling through the regulation of histone modification [243]. miRNA enrichment analysis on the genes of interest in this study yielded the conserved hsa-miR-3135a miRNA as the only significantly enriched miRNA in the HIV positive group associated with disease progression from latent to active TB. This miRNA regulates DESI1 and PRMT1, both of which were associated with H4K20me1 modification. Research has previously identified hsa-miR-3135a as being down-regulated in latent TB compared with active TB [244], but the HIV status of the study group was not reported. This result, and our analysis, suggest that hsa-miR-3135a recruitment to DESI1 and PRMT1 would reduce their expression, resulting in progression from latent to active TB disease. This does appear to be the case with PRMT1, which is downregulated in latent TB and HIV-coinfection in our study. However, the results from our analysis show an inverse correlation between hsa-miR-3135a and DESI1 expression in latent TB and HIV coinfection. Further research is therefore required to clarify the relationship between this miRNA, H4K20me1 and the target genes DESI1 and PRMT1 in the progression from latent to active to TB in HIV infected individuals.

3.4.4 *Challenges in identifying differentially methylated regions from studies using blood samples*

Among all the epigenetic modifications, DNA methylation perturbations have been the most widely studied. Several studies exist describing the reprogramming of DNA methylation patterns in PBMCs after *Mtb* exposure [245]. Due to convenience and ease of sampling, DNA samples used for methylation studies are commonly derived from whole blood as tissue types. However, blood tissue comprises many different cell types in varying proportions and different cell compositions [246]. In addition, DNA methylation profiles show significant variation across tissue types and individual cell types [247]. As such, observed changes in DNA methylation may lead to confounding signals. Over the years, studies have observed that this variation can affect the interpretation of methylation studies based on whole blood [248].

For instance, in the case of latent and active TB, blood samples might be taken from individuals latently infected with *Mtb* and those with active TB disease to look at DNA methylation differences in the two states. The observed changes in DNA methylation at individual CpG sites within genes may be a result of changes in the ratio of different

cell types in latent *Mtb* infection and once the disease has progressed to the active state, and would reflect differences in methylation profiles of different cell types rather than adaptive changes in methylation due to immune response. This could impede the inferences drawn about the functional role of DNA methylation changes in latent and active TB.

We faced constraints in analysing DNA methylation in TB as the transcriptomics dataset used in this study was taken from PBMC samples. The available methylation tracks of the NIH Roadmap Epigenomics Consortium that are of interest to us were those of PMBCs. Since the cellular compositions of those PMBCs would differ, we could not get a sense of differential methylation between the patient groups.

DNA methylation is an important component of the epigenetic landscape; however, the issue lies in how informative DNA methylation patterns in whole blood samples can be. Thus, great caution should be exercised when interpreting methylation profiles from blood samples to draw insights from any differences implicated in a disease. A future avenue for epigenetic studies is the use of cell-free DNA (cfDNA) from blood samples. CfDNA refers to small DNA fragments present in plasma and other body fluids such as urine, cerebral spinal fluid, pleural fluid, saliva, and others [247]. Recent human studies have shown that sequencing of methylation from cfDNA in blood is an accessible and non-invasive method to gain information on the state of various diseases [247]. New studies have even begun exploring *Mtb* cfDNA detection in patients with latent and active TB [249]. This approach could potentially be used to observe DNA in particular tissues for future TB epigenetic studies focusing on HIV infected persons.

3.5 Conclusion

Through the use of TF, HM and miRNA enrichment analyses in specific cell lines of interest, we have identified possible epigenetic marks that underly the regulation of our selected genes of interest involved in latent to active TB disease progression in HIV positive individuals. Following enrichment analyses, an extensive literature search was performed to narrow down candidate epigenetic marks that have potential in unearthing latent TB regulation in HIV positive patients and are recommended for further investigation. Our findings present future opportunities to gather experimental evidence that connects these epigenetic marks during latent TB and HIV coinfection.

4 CONCLUDING REMARKS

4.1 Study Rationale

HIV infection promotes the progression of latent infections of *Mtb* to the active disease with the primary challenge of diagnosis being the development of efficient and sensitive methods to detect latent TB in HIV infected individuals. Previous studies have identified and reported transcriptional signatures for active TB along with signatures predicting the risk of active TB disease in latent TB infected individuals or those with other diseases. Researchers have also identified characteristic genes for active TB in HIV infected patients. However, no studies have identified predictive transcriptional signatures that discriminate latent TB from active TB disease in HIV positive persons. Traditional statistical tests have been widely used for identifying DEGs as biomarkers using microarray gene expression data with the drawback of challenging downstream analysis due to the high dimensionality of the datasets. In recent years, multivariate statistical analyses and machine learning approaches have been developed and applied to microarray datasets. Using an integrative data-driven approach that leverages statistical DE results and results obtained by ML feature selection and classification can provide a viable gene signature that helps understand differences between disease states.

Additionally, ideal biomarkers capable of discriminating between latent TB and active TB in immunocompromised persons need to also be biologically significant and functionally relevant. This requires the use of enrichment analyses to infer networks and associated pathways from expression profiles consequently providing avenues for further investigation into the potential biological mechanisms of sets of genes. Given the pivotal role that epigenetic modulation plays in gene expression, existing studies have explored *Mtb*-induced epigenetic alterations. Although epigenetic regulation appears to be a possible biological factor underlying susceptibility or resistance to latent TB progression to active TB disease in immunocompromised individuals, the mechanisms of these epigenetic modifications are not fully understood. As such, the annotation and retrieval of enriched epigenetic features of a transcriptional gene list can be beneficial for targets in therapies to reduce the activation of TB disease state in HIV infected patients.

Thus, the current study applied a novel data-driven approach that leveraged statistical

differential expression analyses as well as supervised machine learning and feature selection methods to an entire pre-existing transcriptomic dataset and integrated the outcome of the two pipelines to define a latent TB gene signature in HIV infected patients. This was performed in conjunction with enrichment analyses to gain a deeper understanding of the biological networks and pathways the genes of interest are associated with. Our work also used this latent TB transcriptional gene signature to perform epigenetic enrichment analyses to obtain candidate epigenetic marks for latent TB in HIV positive individuals.

4.2 Findings

Extensive computational analyses facilitated the identification of a transcriptomic signature associated with latent TB in HIV positive patients based on samples from an existing microarray dataset. This involved applying ML modelling to the transcriptomic dataset for latent and active TB classification. Overall, the three ML classification models performed extremely well in predicting active TB and latent TB from expression data. The resulting genes from the ML pipeline were integrated with results from a conventional statistical pipeline for DE analysis, namely OPLS-DA and $\log_2$ FC, to define an 11-gene signature for latent TB in HIV positive individuals. We observed numerous genes which appeared more frequently in the different ML classifiers, which suggest functional redundancy in the gene expression patterns that distinguish latent TB from active TB in HIV positive individuals.

These analyses in combination with a literature search enabled us to ascertain the biological functions associated with these genes of interest. Using OPLS-DA and three machine learning approaches, namely logistic regression, support vector classifier and random forest, our study identified a total of 11 DEGs that discriminate between active TB and latent TB in HIV positive patients. A number of these genes are linked to biological processes including the T and B cell receptor signalling pathways, that have been characterised in terms of latent TB and active TB but remain to be characterised in terms of coinfection with HIV. Pathway enrichment analysis on DEGs also revealed several pathways that may be involved in TB and HIV coinfection. These pathways include positive regulation of cellular metabolic process, insulin signalling pathway, endocrine resistance and ATP binding. Although these pathways have not been strongly linked to literature and no genes from our final gene list were associated with them, they require future investigation to establish their possible involvement in latent TB progression in immunocompromised individuals. Additionally, some of the genes

identified in the signature pointed to chromatin regulation, which signified the involvement of epigenetic regulation in the progression from latent to active TB in HIV infected persons.

The enrichment analysis performed allowed for the prediction of epigenetic mechanisms of latent TB and HIV positive associated genes through the identification of TFs and HMs in the defined promoter region of the genes of interest in PBMCs and immune cells. Furthermore, miRNA enrichment analysis on the genes of interest in this study yielded one conserved miRNA as the only significantly enriched miRNA in the HIV positive group associated with disease progression from latent to active TB. While experimental data in HIV infected models are lacking, these enriched TFs, HMs and miRNA identified in PBMCs, B and T cells may provide a focus for future efforts to uncover the latent TB epigenetic landscape in HIV positive individuals and its role in the host immune response.

4.3 Implications of the Study

In this study, we presented a novel approach that leveraged statistical differential expression analyses and supervised ML and feature selection methods to an entire transcriptomic dataset and integrated the outcome of the two pipelines to define a gene signature panel characterising progression from latent to active TB in HIV infected patients. Overall, this work has presented a reliable set of predictive genes that contribute to a better understanding of the biological mechanisms of latent TB in HIV infected persons. This defined gene signature panel was used to ascertain biological pathways, interactions and functions that may be related to latent TB and HIV coinfection. The fact that our study revealed genes that play a significant role in pathways that have been extensively discussed in literature in relation to TB further supports the validity of the approach used to generate the gene signature. The study identified some of the epigenetic alterations associated with latent TB, which could be used as targets in therapies that are aimed are reducing the systematic activation state in HIV infected patients. Using the associated genes for further downstream analyses, our research might fill the gap in understanding the mechanisms of the discussed pathways in the progression of latent TB to active TB, particularly in HIV infected individuals. However, we recommend further studies, especially in sub-Saharan African patients, to confirm the role of these genes in latent TB/HIV coinfection and the associated biological pathways, which would establish their influence in latent TB/HIV coinfection therapies.

4.4 Challenges and Limitations

Some of the challenges faced during this study include the following:

1. One challenge was the possible application of other feature selection methods, which might have been more suitable to the microarray dataset; however, some of these methods are computationally expensive to run on a large expression dataset containing thousands of genes. Although we were restricted in our choice of feature selection methods, we selected the most suitable methods for this type of dataset.
2. Given the limited available data on HIV infected cell lines, during enrichment analysis we selected cell lines involved in TB infection that would be the best substitution (i.e., the major innate immune cell types including PBMCs, T cells, and B cells).
3. We faced constraints in predicting DNA methylation in TB as the transcriptomics dataset used in this study was taken from PBMC samples. The available methylation tracks of the NIH Roadmap Epigenomics Consortium that are of interest to us were those of PMBCs. Since the cellular compositions of those PMBCs would differ, we could not get a sense of differential methylation between the patient groups.

4.5 Future Prospectives

- ML approaches could potentially be used for the prediction of genomic sites that are susceptible to epigenetic modifications [250]. This could significantly increase the potential to develop efficient molecular diagnostics for latent TB in immunocompromised individuals.
- The poor reproducibility of microarray gene expression studies can be overcome by the application of recent meta-analysis approaches [251]. In future, a meta-analysis approach could be taken where existing transcriptome datasets from different studies could be integrated to screen for latent TB biomarkers in patients who are HIV positive.
- There is a need for future studies performing enrichment analyses using experimental data collected from HIV infected cell lines to gain greater clarity on the epigenomic landscape of latent TB/HIV coinfected patients.
- DNA methylation is an important component of the epigenetic landscape; however, the issue lies in how informative DNA methylation patterns in whole blood samples can be. Thus, great caution should be exercised when

interpreting methylation profiles from blood samples to draw insights from any differences implicated in a disease. A future avenue for epigenetic studies is the use of cell-free DNA (cfDNA) from blood samples. Recent human studies have shown that sequencing of methylation from cfDNA in blood is an accessible and non-invasive method to gain information on the state of various diseases [247]. New studies have even begun exploring *Mtb* cfDNA detection in patients with latent and active TB [249]. This approach could potentially be used to observe DNA in particular tissues for future TB epigenetic studies focusing on HIV infected persons.

5 REFERENCES

1. WHO. *Tuberculosis Key Facts* 2020; Available from: https://www.who.int/news-room/fact-sheets/detail/tuberculosis.
2. WHO, *Global Tuberculosis Report 2018* 2018, World Health Organization: Geneva.
3. Lee, S.H., *Tuberculosis Infection and Latent Tuberculosis.* Tuberc Respir Dis (Seoul), 2016. **79**(4): p. 201-206.
4. Young, C. and M. Severn, *Latent Tuberculosis Infection Testing in People with Compromised Immunity Prior to Biologic Therapy: A Review of Diagnostic Accuracy, Clinical Utility, and Guidelines.* Canadian Journal of Health Technologies, 2020: p. 47.
5. Banerjee, U., et al., *Immune Subtyping in Latent Tuberculosis.* Front Immunol, 2021. **12**: p. 595746.
6. Jilani, T.N., et al. *Active Tuberculosis*. 2021 6 August 2021; Available from: https://www.ncbi.nlm.nih.gov/books/NBK513246/.
7. WHO, *Global Tuberculosis Report*. 2020, World Health Organization.
8. UNAIDS, *Tuberculosis and HIV Progress Towards the 2020 Target.* 2019.
9. Teweldemedhin, M., et al., *Tuberculosis-Human Immunodeficiency Virus (HIV) co-infection in Ethiopia: a systematic review and meta-analysis.* BMC Infectious Diseases, 2018. **18**(676): p. 2-9.
10. Martinson, N.A., C.J. Hoffmann, and R.E. Chaisson, *Epidemiology of tuberculosis and HIV: recent advances in understanding and responses.* Proc Am Thorac Soc, 2011. **8**(3): p. 288-93.
11. Michas, F. *Leading causes of death in South Africa 2017, by number of deaths*. 2020; Available from: https://www.statista.com/statistics/1127548/main-causes-of-death-in-south-africa/.
12. Walt, M.v.d. and S. Moyo, *The First National TB Prevalence Survey South Africa*. 2018, The National Institute for Communicable Diseases of South Africa. p. 25.
13. *HIV and AIDS in South Africa* 2020 15 April 2020; Available from: https://www.avert.org/professionals/hiv-around-world/sub-saharan-africa/south-africa.
14. Diaz, A., et al., *Increased Frequency of CD4+ CD25+ FoxP3+ T Regulatory Cells in Pulmonary Tuberculosis Patients Undergoing Specific Treatment and Its Relationship with Their Immune-Endocrine Profile.* Journal of Immunology Research, 2015. **2015**: p. 8.
15. Lee, S.W., et al., *Gene Expression Profiling Identifies Candidate Biomarkers for Active and Latent Tuberculosis* BMC Bioinformatics, 2016 **17**(1): p. 116.
16. Teklu, T., et al., *Potential Immunological Biomarkers for Detection of Mycobacterium tuberculosis Infection in a Setting Where M. tuberculosis Is Endemic, Ethiopia.* Infection and Immunity, 2018 **86**(4): p. 1-11.

17. Roeger, L.I., Z. Feng, and C. Castillo-Chavez, *Modeling TB and HIV co-infections.* Math Biosci Eng, 2009. **6**(4): p. 815-37.
18. Mitku, A., et al., *Prevalence and associated factors of TB/HIV co-infection among HIV Infected patients in Amhara region, Ethiopia.* African Health Sciences 2016. **16**(2): p. 588-595.
19. Ma, Y.-h., et al., *Gene Expression Profile of AIDS Patients with TB Based on Bioinformatics Analysis.* Advances in Social Science, Education and Humanities Research 2018. **170** p. 130-137.
20. Magombedze, G., D. Dowdy, and N. Mulder, *Latent Tuberculosis: Models, Computational Efforts and the Pathogen's Regulatory Mechanisms during Dormancy.* Front Bioeng Biotechnol, 2013. **1**: p. 4.
21. Diedrich, C. and J. Flynn, *HIV-1/Mycobacterium tuberculosis Coinfection Immunology: How Does HIV-1 Exacerbate Tuberculosis?* Infection and Immunity, 2011. **75**(4): p. 1407-1417.
22. Bruchfeld, J., M. Correia-Neves, and G. Kallenius, *Tuberculosis and HIV Coinfection.* Cold Spring Harbor Perspectives in Medicine 2015. **5**(7): p. 1-15.
23. Ugarte-Gil, C., R.M. Carrillo-Larco, and D.E. Kirwan, *Latent tuberculosis infection and non-infectious co-morbidities: Diabetes mellitus type 2, chronic kidney disease and rheumatoid arthritis.* Int J Infect Dis, 2019. **80S**: p. S29-S31.
24. Al-Orainey, I., *Diagnosis of latent tuberculosis: Can we do better?* Thoracic Medicine, 2009. **4**(1): p. 5-9.
25. Banaei, N., R. Gaur, and M. Pai, *Interferon Gamma Release Assays for Latent Tuberculosis: What Are the Sources of Variability?* Journal of Clinical Microbiology, 2016. **54**(4): p. 845-850.
26. Lagrange, P. and J. Herrmann, *Diagnosing Latent Tuberculosis Infection in the HIV Era.* The Open Respiratory Medicine Journal, 2008 **2**: p. 52-59.
27. Yong, Y., et al., *Immune Biomarkers for Diagnosis and Treatment Monitoring of Tuberculosis: Current Developments and Future Prospects.* Frontiers in Microbiology, 2019. **10**(2789): p. 1-18.
28. Dawany, N., et al., *Identification of a 251 Gene Expression Signature That Can Accurately Detect M. tuberculosis in Patients with and without HIV Co-Infection.* PLoS ONE, 2014. **9**(2): p. 8.
29. WHO, *Management of Tuberculosis and HIV Coinfection*. 2013, World Health Organization. p. 1-46.
30. Bares, S.H. and S. Swindells, *Latent Tuberculosis and HIV Infection.* Current Infectious Disease Reports, 2020. **22**(7).
31. Person, A.K. and T.R. Sterling, *Treatment of latent tuberculosis infection in HIV: shorter or longer?* Curr HIV/AIDS Rep, 2012. **9**(3): p. 259-66.
32. *Treatment of LTBI and TB for Persons with HIV*. 2016 Available from: https://www.cdc.gov/tb/topic/treatment/tbhiv.htm.
33. Casamassimi, A., et al., *Transcriptome Profiling in Human Diseases: New Advances and Perspectives.* Int J Mol Sci, 2017. **18**(8).
34. Lowe, R., et al., *Transcriptomics technologies.* PLoS Comput Biol, 2017. **13**(5): p. e1005457.

35. Rao, M.S., et al., *Comparison of RNA-Seq and Microarray Gene Expression Platforms for the Toxicogenomic Evaluation of Liver From Short-Term Rat Toxicity Studies.* Front Genet, 2018. **9**: p. 636.
36. Fajarda, O., et al., *Merging microarray studies to identify a common gene expression signature to several structural heart diseases.* BioData Min, 2020. **13**: p. 8.
37. Anjum, A., et al., *Identification of Differentially Expressed Genes in RNA-seq Data of Arabidopsis thaliana: A Compound Distribution Approach.* J Comput Biol, 2016. **23**(4): p. 239-47.
38. Pham, T., C. Wells, and D. Crane, *Analysis of Microarray Gene Expression Data.* Current Bioinformatics, 2006. **1**(1): p. 37-53.
39. Wang, Z., et al., *Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd.* Nat Commun, 2016. **7**: p. 12846.
40. Singhania, A., et al., *The value of transcriptomics in advancing knowledge of the immune response and diagnosis in tuberculosis.* Nat Immunol, 2018. **19**(11): p. 1159-1168.
41. Kulkarni, V., et al., *A Two-Gene Signature for Tuberculosis Diagnosis in Persons With Advanced HIV.* Front Immunol, 2021. **12**: p. 631165.
42. Ruan, Q.L., et al., *Transcriptional signatures of human peripheral blood mononuclear cells can identify the risk of tuberculosis progression from latent infection among individuals with silicosis.* Emerg Microbes Infect, 2021. **10**(1): p. 1536-1544.
43. Mulenga, H., et al., *Performance of diagnostic and predictive host blood transcriptomic signatures for Tuberculosis disease: A systematic review and meta-analysis.* PLoS One, 2020. **15**(8): p. e0237574.
44. Bayaa, R., et al., *Multi-country evaluation of RISK6, a 6-gene blood transcriptomic signature, for tuberculosis diagnosis and treatment monitoring.* Sci Rep, 2021. **11**(1): p. 13646.
45. Scully, E.P. and B.D. Bryson, *Unlocking the complexity of HIV and Mycobacterium tuberculosis coinfection.* J Clin Invest, 2021. **131**(22).
46. Raza, K., *Application of Data Mining in Bioninformatics.* Indian Journal of Computer Science and Engineering 2010. **1**(2): p. 114-118.
47. Ruskin, H.J. and A. Barat, *Recent advances in computational epigenetics.* Advances in Genomics & Genetics, 2017. **2018**(8): p. 12.
48. Lim, S.J., T.W. Tan, and J.C. Tong, *Computational Epigenetics: the new scientific paradigm.* Biomedical informatics, 2010. **4**(7): p. 6.
49. Duan, Y., et al., *signatureSearch: environment for gene expression signature searching and functional interpretation.* Nucleic Acids Res, 2020. **48**(21): p. e124.
50. Ali, Z. and S.B. Bhaskar, *Basic statistical tools in research and data analysis.* Indian J Anaesth, 2016. **60**(9): p. 662-669.
51. Bandyopadhyay, S., S. Mallik, and A. Mukhopadhyay, *A Survey and Comparative Study of Statistical Tests for Identifying Differential Expression from Microarray Data.* IEEE/ACM Trans Comput Biol Bioinform, 2014. **11**(1): p. 95-115.
52. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic Acids Res, 2015. **43**(7): p. e47.

53. Sherpa, D. *Introduction to Univariate, Bivariate and Multivariate Analysis*. 2021; Available from: https://medium.com/analytics-vidhya/univariate-bivariate-and-multivariate-analysis-8b4fc3d8202c.
54. Dembélé, D. and P. Kastner, *Fold change rank ordering statistics: a new method for detecting differentially expressed genes.* BMC Bioinformatics, 2014. **15**(14): p. 15.
55. Hopkins, S., J.R. Dettori, and J.R. Chapman, *Parametric and Nonparametric Tests in Spine Research: Why Do They Matter?* Global Spine J, 2018. **8**(6): p. 652-654.
56. Boslaugh, S. and P.A. Watters, *Statistics in a Nutshell* First Edition ed. 2008, 1005 Gravenstein Highway North, Sebastopol, CA 95472.: O'Reilly Media, Inc.
57. Piccolo, S.R., et al., 2021.
58. Novianti, P.W., K.C. Roes, and M.J. Eijkemans, *Evaluation of gene expression classification studies: factors associated with classification performance.* PLoS One, 2014. **9**(4): p. e96063.
59. *<12859_2019_3105_MOESM1_ESM.pdf>.*
60. Mahendran, N., et al., *Machine Learning Based Computational Gene Selection Models: A Survey, Performance Evaluation, Open Issues, and Future Research Directions.* Front Genet, 2020. **11**: p. 603808.
61. Heidenreich, H. *What are the types of machine learning?* 2018 5 December 2018; Available from: https://towardsdatascience.com/what-are-the-types-of-machine-learning-e2b9e5d1756f.
62. Zahoor, J. and K. Zafar, *Classification of Microarray Gene Expression Data Using an Infiltration Tactics Optimization (ITO) Algorithm.* Genes (Basel), 2020. **11**(7).
63. Kira, K. and L. Rendell, *A Practical Approach to Feature Selectio.* Machine Learning Proceedings, 1992: p. 7.
64. Halperin, E., G. Kimmel, and R. Shamir, *Tag SNP selection in genotype data for maximizing SNP prediction accuracy.* Bioinformatics, 2005. **21 Suppl 1**: p. i195-203.
65. Sun, L., et al., *Joint neighborhood entropy-based gene selection method with fisher score for tumor classification.* Applied Intelligence, 2019. **49**: p. 14.
66. Pearson, W., et al., *Multi-Round Random Subspace Feature Selection for Incomplete Gene Expression Data.* IEEE, 2019: p. 7.
67. Haynes, W.A., et al., *Differential expression analysis for pathways.* PLoS Comput Biol, 2013. **9**(3): p. e1002967.
68. Signorelli, M., V. Vinciotti, and E.C. Wit, *NEAT: an efficient network enrichment analysis test.* BMC Bioinformatics, 2016. **17**(1): p. 352.
69. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.* Nucleic Acids Res, 2009. **37**(1): p. 1-13.
70. Khatri, P., M. Sirota, and A.J. Butte, *Ten years of pathway analysis: current approaches and outstanding challenges.* PLoS Comput Biol, 2012. **8**(2): p. e1002375.
71. Skinner, M.K., *Endocrine disruptor induction of epigenetic transgenerational inheritance of disease.* Mol Cell Endocrinol, 2014. **398**(1-2): p. 4-12.

72. Nowacka-Zawisza, M. and E. Wisnik, *DNA methylation and histone modifications as epigenetic regulation in prostate cancer (Review).* Oncol Rep, 2017. **38**(5): p. 2587-2596.
73. Okano, M., et al., *DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development.* Cell, 1999. **99**: p. 10.
74. Barau, J., et al., *The DNA methyltransferase DNMT3C protects male germ cells from transposon activity.* Science 2016. **354**: p. 3.
75. Moore, L.D., T. Le, and G. Fan, *DNA methylation and its basic function.* Neuropsychopharmacology, 2013. **38**(1): p. 23-38.
76. Deaton, A.M. and A. Bird, *CpG islands and the regulation of transcription.* Genes Dev, 2011. **25**(10): p. 1010-22.
77. Han, H., et al., *DNA methylation directly silences genes with non-CpG island promoters and establishes a nucleosome occupied promoter.* Hum Mol Genet, 2011. **20**(22): p. 4299-310.
78. Kim, M. and J. Costello, *DNA methylation: an epigenetic mark of cellular memory.* Exp Mol Med, 2017. **49**(4): p. e322.
79. Jin, Z. and Y. Liu, *DNA methylation in human diseases.* Genes Dis, 2018. **5**(1): p. 1-8.
80. Wu, H. and Y. Zhang, *Reversing DNA methylation: mechanisms, genomics, and biological functions.* Cell, 2014. **156**(1-2): p. 45-68.
81. Kohli, R.M. and Y. Zhang, *TET enzymes, TDG and the dynamics of DNA demethylation.* Nature, 2013. **502**(7472): p. 472-9.
82. Phillips, T., *The Role of Methylation in Gene Expression.* Nature Education, 2008. **1**(1): p. 1.
83. Suzuki, M.M. and A. Bird, *DNA methylation landscapes: provocative insights from epigenomics.* Nat Rev Genet, 2008. **9**(6): p. 465-76.
84. Robertson, K.D., *DNA methylation and human disease.* Nature Reviews Genetics, 2005. **6**: p. 13.
85. Das, P.M. and R. Singal, *DNA Methylation and Cancer.* Journal of Clinical Oncology, 2016 **22**(22): p. 10.
86. Murrell, A., P.J. Hurd, and I.C. Wood, *Epigenetic mechanisms in development and disease.* Biochem Soc Trans, 2013. **41**(3): p. 697-9.
87. Zhuang, J., et al., *Perspectives on the Role of Histone Modification in Breast Cancer Progression and the Advanced Technological Tools to Study Epigenetic Determinants of Metastasis.* Front Genet, 2020. **11**: p. 603552.
88. Alaskhar Alhamwe, B., et al., *Histone modifications and their role in epigenetics of atopy and allergic diseases.* Allergy Asthma Clin Immunol, 2018. **14**: p. 39.
89. Harb, H., et al., *Recent developments in epigenetics of pediatric asthma.* Current Opinion in Pediatrics, 2016. **28**(6): p. 9.
90. Swygert, S.G. and C.L. Peterson, *Chromatin dynamics: interplay between remodeling enzymes and histone modifications.* Biochim Biophys Acta, 2014. **1839**(8): p. 728-36.
91. Wapenaar, H. and F.J. Dekker, *Histone acetyltransferases: challenges in targeting bi-substrate enzymes.* Clin Epigenetics, 2016. **8**: p. 59.

92. Morera, L., M. Lubbert, and M. Jung, *Targeting histone methyltransferases and demethylases in clinical trials for cancer therapy.* Clin Epigenetics, 2016. **8**: p. 57.
93. Kaniskan, H.U., M.L. Martini, and J. Jin, *Inhibitors of Protein Methyltransferases and Demethylases.* Chem Rev, 2018. **118**(3): p. 989-1068.
94. Hyun, K., et al., *Writing, erasing and reading histone lysine methylations.* Exp Mol Med, 2017. **49**(4): p. e324.
95. Bannister, A.J. and T. Kouzarides, *Regulation of chromatin by histone modifications.* Cell Res, 2011. **21**(3): p. 381-95.
96. Rossetto, D., N. Avvakumov, and J. Cote, *Histone phosphorylation: a chromatin modification involved in diverse nuclear events.* Epigenetics, 2012. **7**(10): p. 1098-108.
97. Lee, J.S., E. Smith, and A. Shilatifard, *The language of histone crosstalk.* Cell, 2010. **142**(5): p. 682-5.
98. Deng, L., et al., *The role of ubiquitination in tumorigenesis and targeted drug discovery.* Signal Transduct Target Ther, 2020. **5**(1): p. 11.
99. Cao, J. and Q. Yan, *Histone ubiquitination and deubiquitination in transcription, DNA damage response, and cancer.* Front Oncol, 2012. **2**: p. 26.
100. Schwertman, P., S. Bekker-Jensen, and N. Mailand, *Regulation of DNA double-strand break repair by ubiquitin and ubiquitin-like modifiers.* Nature Reviews Molecular Cell Biology volume 2016. **17**: p. 15.
101. Du, J., et al., *DNA methylation pathways and their crosstalk with histone methylation.* Nat Rev Mol Cell Biol, 2015. **16**(9): p. 519-32.
102. Dumitrescu, R.G., *DNA methylation and histone modifications in breast cancer.* Methods in Molecular Biology 2012. **863**: p. 10.
103. Cedar, H. and Y. Bergman, *Linking DNA methylation and histone modification: patterns and paradigms.* Nature Reviews Genetic, 2009. **10**(5): p. 9.
104. Consortium, E.P., et al., *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.* Nature, 2007. **447**(7146): p. 799-816.
105. Chen, L.-L. and G.G. Carmichael, *Decoding the function of nuclear long non-coding RNAs.* Current Opinion in Cell Biology, 2010. **22**(3): p. 7.
106. Xie, W., et al., *Histone h3 lysine 56 acetylation is linked to the core transcriptional network in human embryonic stem cells.* Mol Cell, 2009. **33**(4): p. 417-27.
107. Lujambio, A. and M. Esteller, *How epigenetics can explain human metastasis: a new role for microRNAs.* Cell Cycle, 2009. **8**(3): p. 377-82.
108. Bai, Y., et al., *RNA regulatory networks in animals and plants: a long noncoding RNA perspective.* Brief Funct Genomics, 2015. **14**(2): p. 91-101.
109. Huang, Q., et al., *Mechanistic Insights Into the Interaction Between Transcription Factors and Epigenetic Modifications and the Contribution to the Development of Obesity.* Front Endocrinol (Lausanne), 2018. **9**: p. 370.

110. Kumar, S., et al., *Non-Coding RNAs as Mediators of Epigenetic Changes in Malignancies.* Cancers (Basel), 2020. **12**(12).
111. Alberti, C. and L. Cochella, *A framework for understanding the roles of miRNAs in animal development.* Development 2017. **144**(14): p. 11.
112. Yao, Q., Y. Chen, and X. Zhou, *The roles of microRNAs in epigenetic regulation.* Current Opinion in Chemical Biology, 2019. **51**: p. 6.
113. Ransohoff, J.D., Y. Wei, and P.A. Khavari, *The functions and unique features of long intergenic non-coding RNA.* Nat Rev Mol Cell Biol, 2018. **19**(3): p. 143-157.
114. Maurano, M.T., et al., *Systematic localization of common disease-associated variation in regulatory DNA.* Science, 2012. **337**(6099): p. 1190-5.
115. Melton, C., et al., *Recurrent somatic mutations in regulatory regions of human cancer genomes.* Nat Genet, 2015. **47**(7): p. 710-6.
116. Roadmap Epigenomics, C., et al., *Integrative analysis of 111 reference human epigenomes.* Nature, 2015. **518**(7539): p. 317-30.
117. Yan, X., et al., *Comprehensive Genomic Characterization of Long Non-coding RNAs across Human Cancers.* Cancer Cell, 2015. **28**(4): p. 529-540.
118. S.Latchman, D., *Transcription factors: An overview.* The International Journal of Biochemistry & Cell Biology, 1997. **29**(12): p. 7.
119. Mitsis, T., et al., *Transcription factors and evolution: An integral part of gene expression (Review).* World Academy of Sciences Journal, 2020.
120. Roopra, A., *MAGIC: A tool for predicting transcription factors and cofactors driving gene sets using ENCODE data.* PLoS Comput Biol, 2020. **16**(4): p. e1007800.
121. Rubin, J.D., et al., 2020.
122. Stein, G.S., et al., *Transcription factor-mediated epigenetic regulation of cell growth and phenotype for biological control and cancer.* Adv Enzyme Regul, 2010. **50**(1): p. 160-7.
123. Ahsendorf, T., et al., *Transcription factors, coregulators, and epigenetic marks are linearly correlated and highly redundant.* PLoS One, 2017. **12**(12): p. e0186324.
124. Nightingale, K.P., *Epigenetics – What it is and Why it Matters*, in *Epigenetics for Drug Discovery*. 2016 Royal Society of Chemistry. p. 1-19.
125. Kathirvel, M. and S. Mahadevan, *The Role of Epigenetics in Tuberculosis Infection* Epigenomics, 2016 **10**(2217): p. 13.
126. Ottenhoff, T.H.M., *New pathways of protective and pathological host defense to mycobacteria.* Trends in Microbiology, 2012. **20**(9): p. 9.
127. Azad, A.K., W. Sadee, and L.S. Schlesinger, *Innate immune gene polymorphisms in tuberculosis.* Infect Immun, 2012. **80**(10): p. 3343-59.
128. MonaSingh, VinodYadav, and GobardhanDas, *Host Epigenetic Modifications in Mycobacterium tuberculosis Infection: A Boon or Bane*, in *The Value of BCG and TNF in Autoimmunity* 2018.
129. Milena, S.E., et al., *Epigenetic alterations are associated with monocyte immune dysfunctions in HIV-1 infection.* Scientific Reports 2018. **8**(5505): p. 14.

130. Berry, M.P., et al., *An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis.* Nature, 2010. **466**(7309): p. 973-7.
131. Blankley, S., et al., *The application of transcriptional blood signatures to enhance our understanding of the host response to infection: the example of tuberculosis.* Philos Trans R Soc Lond B Biol Sci, 2014. **369**(1645): p. 20130427.
132. Burel, J.G. and B. Peters, *Discovering transcriptional signatures of disease for diagnosis versus mechanism.* Nat Rev Immunol, 2018. **18**(5): p. 289-290.
133. Zak, D.E., et al., *A blood RNA signature for tuberculosis disease risk: a prospective cohort study.* The Lancet, 2016. **387**(10035): p. 2312-2322.
134. Anderson, S.T., et al., *Diagnosis of childhood tuberculosis and host RNA expression in Africa.* N Engl J Med, 2014. **370**(18): p. 1712-1723.
135. Kaforou, M., et al., *Detection of tuberculosis in HIV-infected and -uninfected African adults using whole blood RNA expression signatures: a case-control study.* PLoS Med, 2013. **10**(10): p. e1001538.
136. Bylesjö, M., et al., *OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification.* Journal of Chemometrics, 2006. **20**(8-10): p. 341-351.
137. Zhang, X., I. Jonassen, and A. Goksoyr, *Machine Learning Approaches for Biomarker Discovery Using Gene Expression Data*, in *Bioinformatics*, I.N. Helder, Editor. 2021: Brisbane (AU).
138. Chen, R.-C., et al., *Selecting critical features for data classification based on machine learning methods.* Journal of Big Data, 2020. **7**(1).
139. Jung, K., T. Friede, and T. Beißbarth, *Reporting FDR analogous confidence intervals for the log fold change of differentially expressed genes.* BMC Bioinformatics, 2011. **12**(288): p. 9.
140. Jafari, M. and N. Ansari-Pour, *Why, When and How to Adjust Your P Values?* Cell J, 2019. **20**(4): p. 604-607.
141. Chakure, A. *Random Forest Classification*. 2019; Available from: https://medium.com/swlh/random-forest-classification-and-its-implementation-d5d840dbead0.
142. Uddin, S., et al., *Comparing different supervised machine learning algorithms for disease prediction.* BMC Med Inform Decis Mak, 2019. **19**(1): p. 281.
143. Kalhori, S.R.N. and X.-J. Zeng, *Evaluation and Comparison of Different Machine Learning Methods to Predict Outcome of Tuberculosis Treatment Course.* Journal of Intelligent Learning Systems and Applications, 2013. **05**(03): p. 184-193.
144. Sauer, C.M., et al., *Feature selection and prediction of treatment failure in tuberculosis.* PLoS One, 2018. **13**(11): p. e0207491.
145. Li, X., et al., *A Debiased MDI Feature Importance Measure for Random Forests.* 2019.
146. Van IJzendoorn, D.G.P., et al., *Machine learning analysis of gene expression data reveals novel diagnostic and prognostic biomarkers and identifies therapeutic targets for soft tissue sarcomas.* PLoS Comput Biol, 2019. **15**(2): p. e1006826.

147. Abbas, M. and Y. El-Manzalawy, *Machine learning based refined differential gene expression analysis of pediatric sepsis.* BMC Med Genomics, 2020. **13**(1): p. 122.
148. Duffy, F.J., et al., *Multinomial modelling of TB/HIV co-infection yields a robust predictive signature and generates hypotheses about the HIV+TB+ disease state.* PLoS One, 2019. **14**(7): p. e0219322.
149. Chen, Y., et al., *Meta-Analysis of Peripheral Blood Transcriptome Datasets Reveals a Biomarker Panel for Tuberculosis in Patients Infected With HIV.* Front Cell Infect Microbiol, 2021. **11**: p. 585919.
150. *Molecular SIgnature Database*. 2005; Available from: http://www2.stat.duke.edu/~sayan/genesets/Jan2006/cards/C3/GGGCGGR_V_SP1_Q6.html.
151. Barthel, R., et al., *Regulation of tumor necrosis factor alpha gene expression by mycobacteria involves the assembly of a unique enhanceosome dependent on the coactivator proteins CBP/p300.* Mol Cell Biol, 2003. **23**(2): p. 526-33.
152. Druszczynska, M., et al., *Monocyte signal transduction receptors in active and latent tuberculosis.* Clin Dev Immunol, 2013. **2013**: p. 851452.
153. Yoo, J.E., et al., *Diabetes Status and Association With Risk of Tuberculosis Among Korean Adults.* JAMA Netw Open, 2021. **4**(9): p. e2126099.
154. Bourgarit, A., et al., *Tuberculosis-associated immune restoration syndrome in HIV-1-infected patients involves tuberculin-specific CD4 Th1 cells and KIR-negative gammadelta T cells.* J Immunol, 2009. **183**(6): p. 3915-23.
155. Guo, H., et al., *Structure of mycobacterial ATP synthase bound to the tuberculosis drug bedaquiline.* Nature. **589**: p. 4.
156. Pawlowski, A., et al., *Tuberculosis and HIV co-infection.* PLoS Pathog, 2012. **8**(2): p. e1002464.
157. Bernal-Fernandez, G., et al., *Decreased expression of T-cell costimulatory molecule CD28 on CD4 and CD8 T cells of mexican patients with pulmonary tuberculosis.* Tuberc Res Treat, 2010. **2010**: p. 517547.
158. Patil, M., *Endocrine and Metabolic Manifestations of Tuberculosis.* General Endocrinology, 2020. **16**(2): p. 8.
159. Silwal, P., et al., *Regulatory Mechanisms of Autophagy-Targeted Antimicrobial Therapeutics Against Mycobacterial Infection.* Front Cell Infect Microbiol, 2021. **11**: p. 633360.
160. Martinez, N., et al., *mTORC2/Akt activation in adipocytes is required for adipose tissue inflammation in tuberculosis.* EBioMedicine, 2019. **45**: p. 314-327.
161. Parandhaman, D.K., L.E. Hanna, and S. Narayanan, *PknE, a serine/threonine protein kinase of Mycobacterium tuberculosis initiates survival crosstalk that also impacts HIV coinfection.* PLoS One, 2014. **9**(1): p. e83541.
162. Harrison, J., et al., *Lcp1 Is a Phosphotransferase Responsible for Ligating Arabinogalactan to Peptidoglycan in Mycobacterium tuberculosis.* mBio, 2016. **7**(4).

163. Nadella, V., et al., *Sphingosine-1-Phosphate (S-1P) Promotes Differentiation of Naive Macrophages and Enhances Protective Immunity Against Mycobacterium tuberculosis.* Front Immunol, 2019. **10**: p. 3085.
164. Li, J., et al., *Mycobacterium tuberculosis Mce3E suppresses host innate immune responses by targeting ERK1/2 signaling.* J Immunol, 2015. **194**(8): p. 3756-67.
165. Duan, L.Y., et al., *Comparative study on the antituberculous effect and mechanism of the traditional Chinese medicines NiuBeiXiaoHe extract and JieHeWan.* Mil Med Res, 2021. **8**(1): p. 34.
166. Dowsett, A., *Anti-VEGF therapies could have role in treating TB.* The Pharmaceutical Journal, 2015. **294**(7848).
167. Datta, M., et al., *Anti-vascular endothelial growth factor treatment normalizes tuberculosis granuloma vasculature and improves small molecule delivery.* Proc Natl Acad Sci U S A, 2015. **112**(6): p. 1827-32.
168. Bouzeyen, R., et al., *FOXO3 Transcription Factor Regulates IL-10 Expression in Mycobacteria-Infected Macrophages, Tuning Their Polarization and the Subsequent Adaptive Immune Response.* Front Immunol, 2019. **10**: p. 2922.
169. Choreno-Parra, J.A., et al., *Thinking Outside the Box: Innate- and B Cell-Memory Responses as Novel Protective Mechanisms Against Tuberculosis.* Front Immunol, 2020. **11**: p. 226.
170. Joosten, S.A., H.A. Fletcher, and T.H.M. Ottenhoff, *A Helicopter Perspective on TB Biomarkers: Pathway and Process Based Analysis of Gene Expression Data Provides New Insight into TB Pathogenesis.* PLoS One. **8**(9): p. 13.
171. Kumar, S.K., et al., *Patterns of T and B cell responses to Mycobacterium tuberculosis membrane-associated antigens and their relationship with disease activity in rheumatoid arthritis patients with latent tuberculosis infection.* PLoS One, 2021. **16**(8): p. e0255639.
172. Butler, R.E., et al., *The balance of apoptotic and necrotic cell death in Mycobacterium tuberculosis infected macrophages is not dependent on bacterial virulence.* PLoS One, 2012. **7**(10): p. e47573.
173. Wong, D., et al., *Protein tyrosine kinase, PtkA, is required for Mycobacterium tuberculosis growth in macrophages.* Sci Rep, 2018. **8**(1): p. 155.
174. Park, J.H., et al., *Understanding Metabolic Regulation Between Host and Pathogens: New Opportunities for the Development of Improved Therapeutic Strategies Against Mycobacterium tuberculosis Infection.* Front Cell Infect Microbiol, 2021. **11**: p. 635335.
175. Ganji, R., et al., *Understanding HIV-Mycobacteria synergism through comparative proteomics of intra-phagosomal mycobacteria during mono- and HIV co-infection.* Sci Rep, 2016. **6**: p. 22060.
176. Lee, S.W., et al., *Gene expression profiling identifies candidate biomarkers for active and latent tuberculosis.* BMC Bioinformatics, 2016. **17 Suppl 1**: p. 3.
177. Sharma, B., et al., *Mycobacterium tuberculosis secretory proteins downregulate T cell activation by interfering with proximal and downstream T cell signalling events.* BMC Immunol, 2015. **16**: p. 67.

178. Mahon, R.N., et al., *Mycobacterium tuberculosis cell wall glycolipids directly inhibit CD4+ T-cell activation by interfering with proximal T-cell-receptor signaling.* Infect Immun, 2009. **77**(10): p. 4574-83.
179. Barham, M.S., et al., *HIV Infection Is Associated With Downregulation of BTLA Expression on Mycobacterium tuberculosis-Specific CD4 T Cells in Active Tuberculosis Disease.* Front Immunol, 2019. **10**: p. 1983.
180. Tran, A. *What's Linear About Logistic Regression*. 2019; Available from: https://towardsdatascience.com/whats-linear-about-logistic-regression-7c879eb806ad.
181. Statnikov, A., L. Wang, and C.F. Aliferis, *A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification.* BMC Bioinformatics, 2008. **9**: p. 319.
182. Mori, Y., et al., *Deep learning-based gene selection in comprehensive gene analysis in pancreatic cancer.* Sci Rep, 2021. **11**(1): p. 16521.
183. Altmann, A., et al., *Permutation importance: a corrected feature importance measure.* Bioinformatics, 2010. **26**(10): p. 1340-7.
184. Chen, B., et al., *Identifying Disease Related Genes by Network Representation and Convolutional Neural Network.* Front Cell Dev Biol, 2021. **9**: p. 629876.
185. Hwang, J.R., et al., *Recent insights of T cell receptor-mediated signaling pathways for T cell activation and development.* Exp Mol Med, 2020. **52**(5): p. 750-761.
186. James, C.A. and C. Seshadri, *T Cell Responses to Mycobacterial Glycolipids: On the Spectrum of "Innateness".* Front Immunol, 2020. **11**: p. 170.
187. Ocana-Guzman, R., et al., *Leukocytes from Patients with Drug-Sensitive and Multidrug-Resistant Tuberculosis Exhibit Distinctive Profiles of Chemokine Receptor Expression and Migration Capacity.* J Immunol Res, 2021. **2021**: p. 6654220.
188. Rijnink, W.F., T.H.M. Ottenhoff, and S.A. Joosten, *B-Cells and Antibodies as Contributors to Effector Immune Responses in Tuberculosis.* Front Immunol, 2021. **12**: p. 640168.
189. Burton, C.T., F.M. Gotch, and N. Imami, *CCR2/64I mutation detection in a HIV-1-positive patient with slow CD4 T-cell decline and delay in disease progression.* International Journal of STD & AIDS, 2005. **16**(5): p. 3.
190. Wachira, D., et al., *Chemokine Coreceptor-2 Gene Polymorphisms among HIV-1 Infected Individuals in Kenya.* Dis Markers, 2015. **2015**: p. 952067.
191. Long, T., et al., *Identification of differentially expressed genes and enriched pathways in lung cancer using bioinformatics analysis.* Mol Med Rep, 2019. **19**(3): p. 2029-2040.
192. Maglione, P.J., et al., *Fc gamma receptors regulate immune activation and susceptibility during Mycobacterium tuberculosis infection.* J Immunol, 2008. **180**(5): p. 3329-38.
193. Pisu, D., et al., *Single cell analysis of M. tuberculosis phenotype and macrophage lineages in the infected lung.* J Exp Med, 2021. **218**(9).

194. Kassa, D., et al., *Discriminative expression of whole blood genes in HIV patients with latent and active TB in Ethiopia.* Tuberculosis (Edinb), 2016. **100**: p. 25-31.
195. Chan, J., et al., *The role of B cells and humoral immunity in Mycobacterium tuberculosis infection.* Semin Immunol, 2014. **26**(6): p. 588-600.
196. Lamptey, H., et al., *Are Fc Gamma Receptor Polymorphisms Important in HIV-1 Infection Outcomes and Latent Reservoir Size?* Front Immunol, 2021. **12**: p. 656894.
197. Bedsaul, J.R., et al., *Mechanisms of Regulated and Dysregulated CARD11 Signaling in Adaptive Immunity and Disease.* Front Immunol, 2018. **9**: p. 2105.
198. Bai, X., et al., *Inhibition of nuclear factor-kappa B activation decreases survival of Mycobacterium tuberculosis in human macrophages.* PLoS One, 2013. **8**(4): p. e61925.
199. *CARD11 gene*. 2020 [cited 2021 23 October 2021]; Available from: https://medlineplus.gov/genetics/gene/card11/.
200. de Carvalho, P.S., F.E. Leal, and M.A. Soares, *Clinical and Molecular Properties of Human Immunodeficiency Virus-Related Diffuse Large B-Cell Lymphoma.* Front Oncol, 2021. **11**: p. 675353.
201. Wong, L.M. and G. Jiang, *NF-kappaB sub-pathways and HIV cure: A revisit.* EBioMedicine, 2021. **63**: p. 103159.
202. Yang, J., et al., *Molecular characterization of T cell receptor beta variable in the peripheral blood T cell repertoire in subjects with active tuberculosis or latent tuberculosis infection.* BMC Infectious Diseases, 2013. **13**(423): p. 10.
203. Vrieling, F., et al., *Analyzing the impact of Mycobacterium tuberculosis infection on primary human macrophages by combined exploratory and targeted metabolomics.* Sci Rep, 2020. **10**(1): p. 7085.
204. Sáez-Cirión, A. and I. Sereti, *Immunometabolism and HIV-1 pathogenesis: food for thought.* Nature Reviews Immunology, 2020. **21**: p. 14.
205. Pedro, M.N., et al., *Insulin Resistance in HIV-Patients: Causes and Consequences.* Front Endocrinol (Lausanne), 2018. **9**: p. 514.
206. Hruz, P.W., *Molecular mechanisms for insulin resistance in treated HIV-infection.* Best Pract Res Clin Endocrinol Metab, 2011. **25**(3): p. 459-68.
207. Patil, M., *Endocrine and Metabolic Manifestations of Tuberculosis.* General Endocrinology, 2020. **16**(2): p. 9.
208. Adamu, B. and S.-B. Fatim, *Endocrine Manifestations of HIV Infection*, in *Current Perspectives in HIV Infection*. 2013.
209. Kalra, S., H. Sleim, and N. Kotwal, *Human immunodeficiency virus and the endocrine system.* Indian J Endocrinol Metab, 2011. **15**(4): p. 231-3.
210. Guo, H., et al., *Structure of mycobacterial ATP synthase bound to the tuberculosis drug bedaquiline.* Nature, 2021. **589**(7840): p. 143-147.
211. Jennelle, L., et al., *HIV-1 protein Nef inhibits activity of ATP-binding cassette transporter A1 by targeting endoplasmic reticulum chaperone calnexin.* J Biol Chem, 2014. **289**(42): p. 28870-84.

212. Stein, C.M., et al., *Genomics of human pulmonary tuberculosis: from genes to pathways.* Curr Genet Med Rep, 2017. **5**(4): p. 149-166.
213. Shim, S. and P.J. Seo, *EAT-UpTF: Enrichment Analysis Tool for Upstream Transcription Factors of a Group of Plant Genes.* Frontiers in Genetics, 2020. **11**: p. 8.
214. Tipney, H. and L. Hunter, *An introduction to effective use of enrichment analysis software.* Human Genomics, 2010. **4**: p. 5.
215. Huang, D.W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nature Protocols, 2009. **4**(1): p. 13.
216. Meiler, A., et al., *AnnoMiner is a new web-tool to integrate epigenetics, transcription factor occupancy and transcriptomics data to predict transcriptional regulators.* Sci Rep, 2021. **11**(1): p. 15463.
217. Riffo-Campos, A.L., I. Riquelme, and P. Brebi-Mieville, *Tools for Sequence-Based miRNA Target Prediction: What to Choose?* Int J Mol Sci, 2016. **17**(12).
218. Cai, Y., et al., *A Brief Review on the Mechanisms of miRNA Regulation.* Genomics, Proteomics & Bioinformatics, 2009. **7**(4): p. 147-154.
219. Seitz, H., *Redefining microRNA targets.* Curr Biol, 2009. **19**(10): p. 870-3.
220. Licursi, V., et al., *MIENTURNET: an interactive web tool for microRNA-target enrichment and network-based analysis.* BMC Bioinformatics, 2019. **20**(1): p. 545.
221. Georgakilas, G.K., N. Perdikopanis, and A. Hatzigeorgiou, *Solving the transcription start site identification problem with ADAPT-CAGE: a Machine Learning algorithm for the analysis of CAGE data.* Sci Rep, 2020. **10**(1): p. 877.
222. JF, R.-S., et al., *A novel role of Yin-Yang-1 in pulmonary tuberculosis through the regulation of the chemokine CCL4.* Tuberculosis, 2016. **96**: p. 8.
223. Zhang, W., et al., *Mycobacterium tuberculosis H37Rv infection regulates alternative splicing in Macrophages.* Bioengineered, 2018. **9**(1): p. 203-208.
224. Kumar, M., et al., *Activating transcription factor 3 modulates the macrophage immune response to Mycobacterium tuberculosis infection via reciprocal regulation of inflammatory genes and lipid body formation.* Cell Microbiol, 2020. **22**(3): p. e13142.
225. Cavalcanti, F., et al., *Efficacy of vafidemstat in experimental autoimmune encephalomyelitis highlights the KDM1A/RCOR1/HDAC epigenetic axis in multiple sclerosis.* Research Square, 2020: p. 45.
226. Singh, V., et al., *Histone Methyltransferase SET8 Epigenetically Reprograms Host Immune Responses to Assist Mycobacterial Survival.* J Infect Dis, 2017. **216**(4): p. 477-488.
227. Fatima, S., et al., *Epigenetic code during mycobacterial infections: therapeutic implications for tuberculosis.* FEBS J, 2021.
228. Qin, W., B.P. Scicluna, and T. van der Poll, *The Role of Host Cell DNA Methylation in the Immune Response to Bacterial Infection.* Front Immunol, 2021. **12**: p. 696280.

229. Verheul, T.C.J., et al., *The Why of YY1: Mechanisms of Transcriptional Regulation by Yin Yang 1.* Front Cell Dev Biol, 2020. **8**: p. 592164.
230. Ganguli, G., U. Mukherjee, and A. Sonawane, *Peroxisomes and Oxidative Stress: Their Implications in the Modulation of Cellular Immunity During Mycobacterial Infection.* Front Microbiol, 2019. **10**: p. 1121.
231. Hogan, D. and R.T. Wheeler, *The complex roles of NADPH oxidases in fungal infection.* Cell Microbiol, 2014. **16**(8): p. 1156-67.
232. Margaritis, M., *Endothelial dysfunction in HIV infection: experimental and clinical evidence on the role of oxidative stress.* Annals of Research Hospitals, 2019. **3**: p. 7-7.
233. Jiang, Y., et al., *The role of HIV Tat protein in HIV-related cardiovascular diseases.* J Transl Med, 2018. **16**(1): p. 121.
234. Wu, R.F., et al., *HIV-1 Tat activates dual Nox pathways leading to independent activation of ERK and JNK MAP kinases.* Journal of Biological Chemistry, 2007. **282**(52): p. 7.
235. Raja, R., et al., *Serum deprivation/starvation leads to reactivation of HIV-1 in latently infected monocytes via activating ERK/JNK pathway.* Sci Rep, 2018. **8**(1): p. 14496.
236. Do, D.V., et al., *SRSF3 maintains transcriptome integrity in oocytes by regulation of alternative splicing and transposable elements.* Cell Discov, 2018. **4**: p. 33.
237. Paz, S., A.R. Krainer, and M. Caputi, *HIV-1 transcription is regulated by splicing factor SRSF1.* Nucleic Acids Res, 2014. **42**(22): p. 13812-23.
238. Ismail, T., et al., *KDM1A microenvironment, its oncogenic potential, and therapeutic significance.* Epigenetics Chromatin, 2018. **11**(1): p. 33.
239. Le Douce, V., et al., *LSD1 cooperates with CTIP2 to promote HIV-1 transcriptional silencing.* Nucleic Acids Res, 2012. **40**(5): p. 1904-15.
240. Wang, Z., et al., *Combinatorial patterns of histone acetylations and methylations in the human genome.* Nat Genet, 2008. **40**(7): p. 897-903.
241. Yaseen, I., et al., *Mycobacteria modulate host epigenetic machinery by Rv1988 methylation of a non-tail arginine of histone H3.* Nat Commun, 2015. **6**: p. 8922.
242. Sharma, G., et al., *Genome-wide non-CpG methylation of the host genome during M. tuberculosis infection.* Sci Rep, 2016. **6**: p. 25006.
243. Wei, J.W., et al., *Non-coding RNAs as regulators in epigenetics (Review).* Oncol Rep, 2017. **37**(1): p. 3-9.
244. Fu, Y., et al., *Deregulated microRNAs in CD4+ T cells from individuals with latent tuberculosis versus active tuberculosis.* J Cell Mol Med, 2014. **18**(3): p. 503-13.
245. Karlsson, L., et al., *A differential DNA methylome signature of pulmonary immune cells from individuals converting to latent tuberculosis infection.* Sci Rep, 2021. **11**(1): p. 19418.
246. Husby, A., *On the Use of Blood Samples for Measuring DNA Methylation in Ecological Epigenetic Studies.* Integr Comp Biol, 2020. **60**(6): p. 1558-1566.

247. Houseman, E.A., et al., *DNA Methylation in Whole Blood: Uses and Challenges.* Curr Environ Health Rep, 2015. **2**(2): p. 145-54.
248. Reinius, L.E., et al., *Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility.* PLoS One, 2012. **7**(7): p. e41361.
249. Pan, S.W., et al., *Mycobacterium tuberculosis-derived circulating cell-free DNA in patients with pulmonary tuberculosis and persons with latent tuberculosis infection.* PLoS One, 2021. **16**(6): p. e0253879.
250. Holder, L.B., M.M. Haque, and M.K. Skinner, *Machine learning for epigenetics and future medical applications.* Epigenetics, 2017. **12**(7): p. 505-514.
251. Tseng, G.C., D. Ghosh, and E. Feingold, *Comprehensive literature review and statistical considerations for microarray meta-analysis.* Nucleic Acids Res, 2012. **40**(9): p. 3785-99.

6 SUPPLEMENTARY INFORMATION

Figure S1.
Signalling network constructed using the gene set of the HIV positive group which occurred in the OPLS-DA model.
(JPEG)

Figure S2.
Signalling network constructed using the gene set of the HIV positive group which occurred in the RF model.
(JPEG)

Figure S3.
Signalling network constructed using the gene set of the HIV positive group which occurred in the LR model.
(JPEG)

Figure S4.
Signalling network constructed using the gene set of the HIV positive group which occurred in the SVC model.
(JPEG)

Table S1.
Microarray Dataset used for Statistical Analyses.
(XLSX)

Table S2.
Cumulative frequencies of the genes that appear in the Top 10 across 5000 iterations of the feature selection process following ML modelling with pre-filtering at FDR < 0.05.
(XLSX)

Table S3.
Final list of top upregulated and downregulated DEGs in HIV positive and all patient groups.
(XLSX)

Table S4.
Enriched network pathway terms that were identified using the OPLS-DA and three ML models in HIV positive patients using the combined KEGG and gene ontology databases.
(XLSX)

Table S5.
Transcription factor enrichment analysis results.
(XLSX)

Table S6.
Histone modification enrichment analysis results.
(XLSX)

Table S7.
miRNA enrichment analysis results.
(XLSX)

Notebook S1.
OPLS-DA modelling in Python (example)
(IPYNB)

Notebook S2.
FDR procedure and log2 fold change calculation (example)
(IPYNB)

Notebook S3.
Selection of Top 10 genes by ML and feature selection (example)
(IPYNB)

Workbook S1.
Comparison of the performance of the different ML classifiers in the two patient groups. Z-score frequencies of the top genes identified using the average log2FC and the OPLS-DA standardised regression coefficients (in HIV positive and all patients at different FDR values.
(TWBX)

Workbook S2.
Pearson correlation matrix created in Tableau of ML classifiers in the different groups at different FDR values.
(TWBX)

Workbook S3.
Cumulative frequencies of genes in the Top 10 by ML algorithm, group and pre-filter method.
(TWBX)